1989

Analysis of one-way layout of count data.

Rajesh Kumar. Barnwal
University of Windsor

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

ANALYSIS OF ONE-WAY LAYOUT OF COUNT DATA

BY

RAJESH KUMAR BARNWAL

A Dissertation
submitted to the
Faculty of Graduate Studies and Research
through the Department of
Mathematics and Statistics in Partial Fulfillment
of the requirements for the Degree
of Doctor of Philosophy at
the University of Windsor

Windsor, Ontario, Canada
1989

ISBN 0-315-61019-0

ABSTRACT

Inadequacy of the Poisson assumption, due to the presence of overdispersion, in analysing count data has been reported by several authors (see McCaughran and Arnold (1976), Bliss and Owen(1958) etc.). Negative binomial distribution has been widely used to incorporate overdispersion in analysing the count data. Several test statistics for detecting negative binomial variation have been presented . $C(\alpha)$ tests, range-justified tests (appealing to the nonnegativity of the dispersion parameter) are compared with the statistic presented by Collings and Margolin(1985).

One-way layout of data in the form of counts is often reported as a result of laboratory experiment or field work. Assuming the underlying distribution for the groups to be negative binomial with common dispersion parameter, two $C(\alpha)$ tests are developed for comparing the means of the groups. Their performance is compared in terms of level and power with the likelihood ratio test and test based on variance stabilising transformation(Anscombe(1948)). A test for checking the validity of assumption of common dispersion is also developed.

In several situations the assumption of a common dispersion parameter might not be tenable. A $C(\alpha)$ test is derived for comparing the means of negative binomial

distributions with unequal dispersion parameters. For two groups, this test is compared with the Welch's approximate degree of freedom formula and Banerji's procedure(1960) for empirical level and power.

Methods for testing the presence of an outlier in data coming from a population following Poisson distribution have also been derived .

In deriving the C(α) test statistics for the above problems, the method presented by Neyman(1959)) has been presented under a more general setting, which covers many situations concerning inferences on several parameters in presence of nuisance parameters.

Respectfully Dedicated to my parents

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Professor S.R. Paul for suggesting the topic of this dissertation and for the advice, help and encouragement that he has given generously.

I would like to thank Dr. D.J. Britten and Dr. R.M. Barron for providing me with the graduate assistantships and Dr. S. R. Paul for research assistantships through his NSERC grant no. 9570, throughout my graduate studies.

I would like to thank the Faculty of Graduate studies and Research for offering the university of Windsor Tuition scholarships .

I would like to thank my teachers Dr. D.S. Tracy, Dr. M.M. Shoukri, Dr. K.Y. Fung and Dr. Tim Traynor for their readiness to help and stimulating discussions during course work and after.

I would like to thank my friends Dr. I.U.H. Tajuddin, Dr. M. Tang, Dr. R. Sharma, Dr. M.H. Hamdan, Dr. S. Kasiviswanathan, Mrs. Shagufta A. Sultan, Mr. K.A. Khan, Mr. I.U.H. Mian, Ms. F. Labropulu, Mr. I. Hussain, Mr. P. Nguyen, Ms. K. Thiagarajah, Mr. W. Sun, Mr. T. Wang, Mrs. W. Obuchowska, Mr. S. Zhang and Ms. I. Ten-Elshof for their support, encouragement and affection .

LIST OF TABLES

TABLE OF CONTENTS

Page

CHAPTER 4

ANALYSIS OF ONE-WAY LAYOUT OF COUNT DATA WITH NEGATIVE BINOMIAL VARIATION WHEN THE DISPERSION PARAMETERS ARE EQUAL

CHAPTER 5

ANALYSIS OF ONE-WAY LAY-OUT OF COUNT DATA WITH NEGATIVE BINOMIAL VARIATION WHEN THE DISPERSION PARAMETERS ARE UNEQUAL

CHAPTER 6

DETECTION OF OUTLIERS IN POISSON SAMPLES

CHAPTER 7

CONCLUSIONS

CHAPTER 1

INTRODUCTION

Much of the methodology concerning analysis of count data is based on the assumption that the data are distributed according to the Poisson distribution. Paul and Plackett(1978) summarise the usual assumptions as

i) the expected values are linear in logarithmic scale,

ii) observations are independently distributed according to Poisson distribution.

The analysis of contingency tables has so far been performed assuming that cell counts follow Poisson distribution (see Fienberg, Bishop & Holland(1975), Plackett (1981)). However, in many practical situations count data exhibit over-dispersion. Poisson distrbution has equal mean and variance, hence over-dispersion or extra-poisson variation is suspected when variance is larger than the mean. One possible relationship between mean and variance considered by Bartlett(1936) and Paul and Plackett(1978) is

$$\sigma^2 = c_1 m + c_2 m^2. \qquad (1.1)$$

When $c_1 = 1$ and $c_2 = c(> 0)$, we get

$$\sigma^2 = m + cm^2,$$

which is the variance of a negative binomial distribution with parameters m and c(see Collings(1981)). Negative binomial distribution is also described as mixed Poisson distribution. There are other distributions in the literature which also model extra-poisson variation like the Sichel's distribution, logarithmic distribution, Polya-Aeppli distribution, etc. The convenience and tractability of the negative binomial distribution provides an attaractive alternative to the Poisson distribution, when data show an evidence of over-dispersion. Numerous examples in biostatistics have shown the inadequacy of the Poisson model and suitability of negative binomial distribution in describing data in the form of counts (see Bliss and Fisher(1953), Bliss and Owen(1958), McCaughran and Arnold(1976), Collings and Margolin(1985)). After comparing several other distributions exhibiting over-dispersion. Anscombe(1950) suggests that negative binomial distribution should be used over others.
The relative ease of calculation of maximum likelihood estimates and the presence of single mode weigh heavily in favour of choice of negative binomial distribution for describing over-dispersion. Different authors have expressed the negative binomial distribution in different forms; see for example Bliss and Fisher (1953), Johnson and Kotz(1969), Bliss and Owen(1958), and Collings and Margolin(1985). The most common form derived as the

probability of getting Y failures before obtaining k successes is given by

$$P(Y = y) = \binom{y+k-1}{k-1} p^k (1-p)^y, \qquad (1.2)$$

where

p is the probability of success at a trial.

Hinz & Gurland(1968,1970) discuss the estimation of k and p in detail. Further, Wilson et al.(1984) discuss the MSE and bias of the maximum likelihood and method of moments estimates of k and p. However, this form is not convenient for comparing the means of several negative binomial distributions.

Anscombe(1950) suggests another form with mean m and exponent k given by

$$P(Y = y) = \frac{\Gamma(y+k)}{y!\Gamma(k)} \left(\frac{m}{m+k}\right)^y \left(\frac{k}{m+k}\right)^k, \qquad (1.3)$$

where

$y = 0,1,2,\ldots$ and $m,k > 0$.

As $k \longrightarrow \infty$, the negative binomial distribution converges to Poisson distribution. By using the reparametrisation $k = c^{-1}$ in (1.3), a more convenient form of the negative binomial distribution was introduced by Bliss and Owen(1958) and used by Collings(1981) and Collings and Margolin (1985) in which the random variable X follows a negative binomial distribution with mean m and dispersion parameter c, denoted by $X \sim NB(m,c)$, if the

probability mass function is given by

$$P(X = x) = \frac{\Gamma(x + c^{-1})}{\Gamma(c^{-1})\, x!} \left[\frac{cm}{1 + cm}\right]^{x} \left[\frac{1}{1 + cm}\right]^{c^{-1}}, \quad (1.4)$$

for $x = 0,1,2,\ldots,$

$m > 0,$

and $c > 0.$

Here $E(X) = m$ and $Var(X) = m + cm^2$. Evidently, the $NB(m,c)$ becomes the Poisson distribution in the limit when $c \to 0$. Further properties of $NB(m,c)$ are given by Paul & Plackett (1978), Anscombe(1950), etc .

The main purpose of this thesis is to analyse one-way lay-out of count data with over-dispersion described by negative binomial distribution. This thesis is also concerned with detection of outliers in a single Poisson sample.

In chapter 2, we review the theory of $C(\alpha)$ tests (Neyman, 1959) and develop test-statistics for three commonly occurring problems in Statistics. The LR test and Wald's test are discussed and compared with the $C(\alpha)$ test.

In chapter 3, we discuss procedures for testing equality of k Poisson distributions. Several test statistics for detecting negative binomial over-dispersion are developed and discussed. A small scale simulation study is conducted

to compare the empirical size and power of these statistics.

In chapter 4, we develop test statistics for testing equality of means of several groups of count data in presence of common dispersion parameter. Two C(α) tests, a likelihood ratio test and two more statistics based on transformed data (Anscombe, 1948) are developed and then compared in terms of size and power using Monte Carlo simulations. The C(α) statistics for testing the equality of the dispersion parameters of several groups of count data are also derived.

In chapter 5, the assumption of common c is dropped . A C(α) statistic is derived , from which two other procedures are also developed . The three procedures are then compared for two groups in terms of size and power.

Chapter 6 deals with detection of a single outlier in a Poisson sample. Several methods are presented and their performances in terms of size and power are compared.

CHAPTER 2

C(α) TESTS AND OTHER ASYMPTOTICALLY OPTIMAL TESTS

2.1 Introduction

In this chapter,it is attempted to describe the C(α) tests in a more general setting than presented by Neyman(1959), and used by Neyman and Scott(1966), Moran (1970), Subrahmaniam(1966), Kocherlakota & Kocherlakota (1985) and others. Its relation to the usual score test and Wald statistic is also discussed. It is shown that the C(α) test proposed by Neyman provides an elegant method of construction of tests for composite hypotheses. Several situations, commonly occurring in data analysis, have been presented in a way which can readily be used in constructing asymptotically optimal test criteria. It has been attempted to provide a unified treatment of tests based on functions of likelihood. The likelihood ratio test requires the convergence of log of density function, while the C(α) test requires the convergence of efficient scores or first partial derivative of logarithm of likelihood. The first two sections explain the construction of C(α) tests. The next three sections present three common problems in Statistics :

i) testing of homogeneity of certain parameter in the presence of nuisance parameters ,

ii) testing of an outlier

and

iii) testing of a set of linear constraints .

In the remaining sections , Wald's test and likelihhod ratio tests and their relation to the C(α) tests are discussed.

2.2 Description of the problem:

Let $(\Omega, \mathcal{A}, P)$ be a probability space and X a real valued measurable function defined on Ω. The probability space induced by X is $(\mathcal{R}, \mathcal{B}, P_X)$. Let P_X represent a density function $f(\theta, x)$,

where

$$\theta = (\theta_1, \theta_2) \in \Theta_{11} \times \Theta_{12} = \Theta_1$$

Our objective is to construct tests for the composite hypothesis given by

H_0: $\theta \in \Theta_0$ $(\subset \Theta_1)$

against

H_A: $\theta \notin \Theta_0$. (2.2.1)

In the absence of any other information on the density function or the subset of the parameter space, we construct the likelihood ratio test. Let ℓ_0 and ℓ_1 be the log-likelihood under the null and the alternative hypotheses respectively and $\hat{\ell}_0$ and $\hat{\ell}_1$ be the corresponding maxima.

Then the statistic

$\chi_1^2 = 2\ (\hat{\ell}_1 - \hat{\ell}_0\ ) \sim \chi^2$-distribution with suitable d.f.

Consider a transformation T,

$$T : \Theta_{11} \times \Theta_{12} \longrightarrow \Theta_{21} \times \Theta_{22} ,$$

i.e. $T(\theta_1,\theta_2) = (\xi_1,\xi_2)$ (2.2.2)

such that whenever $\xi_1=\xi_{10}$ (a fixed quantity) then

$$T^{-1}((\xi_{10},\xi_2) : \xi_2 \in \Theta_{22}) = \Theta_o .$$

If T is 1-1 and onto then the testing problem described in (2.2.1) reduces to

$$H_o : \xi_1 = \xi_{10} , \xi_2 \in \Theta_{22}$$

and

$$H_A : \xi_1 \neq \xi_{10} , \xi_2 \in \Theta_{22} . \quad (2.2.3)$$

Let the new density function be $g((\xi_1,\xi_2),x)$. If the composite hypotheses are already presented in the form (2.2.3), then we do not look for the transformation T, and proceed to construct C(α) test. Another likelihood ratio test in terms of (ξ_1,ξ_2) can be constructed for the hypotheses in (2.2.1), in some cases it may result in a simpler form.

Let $\chi_n(\xi,x) = (x_1,x_2,\ldots,x_n)$ be a sample of n observations from a population with density $g(\xi,x)$. Let W_n be the sample space and ω_n^o be the critical region for (2.2.3) associated with $\chi_n(\xi,x)$.

Definition 1 : If a sequence of critical regions $\{\omega_n\}$ has the property that for all $\xi_2 \in \Theta_{22}$,

$$\lim_{n\to\infty} P\{ \chi_n\{\xi_{10},\xi_2\} \in \omega_n\} = \alpha , \quad (2.2.4)$$

then we say that $\{\omega_n\}$ is an asymptotic test of the hypothesis H_o at level of significance α.

Let $K(\alpha)$ be a class of asymptotic tests of the hypothesis H_o ,all corresponding to the same level α. Let $\xi^* = \{\xi_{1n}\}$ be a sequence of points belonging to Θ_{21} and converging to ξ_{10}. Let Γ denote a certain class of sequences ξ^* and let $\{\omega_n^o\} \in K(\alpha)$.

Definition 2: With reference to Γ, the test $\{\omega_n^o\}$ is optimal within the class $K(\alpha)$ if for all sequences of $\{\omega_n\} \in K(\alpha)$ and sequence $\xi^* \in \Gamma$ and $\xi_2 \in \Theta_{22}$, when the following holds

$$\lim_{n\to\infty} \left[P\left[\chi_n(\xi_{1n},\xi_2) \in \omega_n^o \right] - P\left[\chi_n(\xi_{1n},\xi_2) \in \omega_n \right] \right] \geq 0 . \qquad (2.2.5)$$

This test may be described as Locally Asymptotically Most Powerful (LAMP) test. Neyman(1959) starts with the hypothesis (2.2.3) ,but (2.2.3) can be arrived at through a transformation T ,if it exists, for a more general hypothesis also.

2.3 Construction of $C(\alpha)$ tests

Let $X_1, X_2, .., X_n$ be observations from the population with density function $g((\xi_1,\xi_2),x)$ where $\xi_1 \in \mathcal{R}^s$ and $\xi_2 \in \mathcal{R}^{p-s}$ and let ℓ be the log-likelihood of the data given by

$$\ell = \sum_{i=1}^{n} \ln(g((\xi_1,\xi_2),x)) \quad . \qquad (2.3.1)$$

Define

$$\frac{\partial \ell}{\partial \xi_1} = \left[\frac{\partial \ell}{\partial \xi_{11}}, \frac{\partial \ell}{\partial \xi_{12}}, \ldots, \frac{\partial \ell}{\partial \xi_{1s}} \right]' \Bigg|_{\xi_1 = \xi_{10}}$$

and (2.3.2)

$$\frac{\partial \ell}{\partial \xi_2} = \left[\frac{\partial \ell}{\partial \xi_{21}}, \frac{\partial \ell}{\partial \xi_{22}}, \ldots, \frac{\partial \ell}{\partial \xi_{2,p-s}} \right]' \Bigg|_{\xi_1 = \xi_{10}} .$$

These partial derivatives of log-likelihood are assumed to be Cramer functions. Cramer(1946) has shown that under the null hypothesis($\frac{\partial \ell}{\partial \xi_1'}, \frac{\partial \ell}{\partial \xi_2'}$)' follows a multivariate normal distribution with mean vector 0 and

dispersion matrix $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$,

where

$$\Sigma_{11} = - E \left[\frac{\partial^2 \ell}{\partial \xi_1 \partial \xi_1'} \Big| \xi_1 = \xi_{10} \right] ,$$

$$\Sigma_{12} = - E \left[\frac{\partial^2 \ell}{\partial \xi_1 \partial \xi_2'} \Big| \xi_1 = \xi_{10} \right] \qquad (2.3.4)$$

and

$$\Sigma_{22} = - E \left[\frac{\partial^2 \ell}{\partial \xi_2 \partial \xi_2'} \Big| \xi_1 = \xi_{10} \right] .$$

Define $S = \frac{\partial \ell}{\partial \xi_1} - B \frac{\partial \ell}{\partial \xi_2}$, where B is the partial regression coefficients matrix obtained by regressing

$\frac{\partial \ell}{\partial \xi_1}$ on $\frac{\partial \ell}{\partial \xi_2}$.

From Anderson(1986) ,

$$B = \Sigma_{12}\Sigma_{22}^{-1} ,$$

Dispersion matrix of S is $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Sigma_{11.2}$

and

$$S'\Sigma_{11.2}^{-1}S \sim \chi^2(s). \qquad (2.3.5)$$

But this expression involves (p-s) unknown nuisance parameters, which makes the statistic unsuitable for actually deciding whether or not to reject the hypothesis. Assuming n to be large, we replace the unknown parameters by their $\sqrt{n}$-consistent estimates obtained from the data. Hence following Neyman(1959)

$$\chi^2_{C\alpha} = \hat{S}'\hat{\Sigma}_{11.2}^{-1}\hat{S} \sim \chi^2(s) . \qquad (2.3.6)$$

$\chi^2_{C\alpha}$ in (2.3.6) reduces to the test statistic obtained by Kocherlakota and Kocherlakota(1985), when s = 1, i.e., ($\xi_1 \in \mathcal{R}$ or $\mathcal{R}_+$) .

2.4 $\sqrt{n}$ - consistent estimators

In the setting decribed earlier, if there exist numbers $A_{ij} \neq 0$ s.t. as $n \to \infty$, the product

$$\left| \hat{\xi}_{2i,n} - \xi_{2i} - A_{ij}(\xi_{1,j} - \xi_{10,j}) \right| \sqrt{n} \qquad (2.4.1)$$

remains bounded in probability for all (ξ_1, ξ_2), then we say that $\{\hat{\xi}_{2n}\}$ forms a locally $\sqrt{n}$- consistent estimate of ξ_2. In case, $A_{ij} = 0$ for all i and j , the sequence $\{\hat{\xi}_{2n}\}$ is called $\sqrt{n}$ - consistent estimator.

Theorem 2.4.1 : Let $\{\hat{\theta}_n\}$ be a sequence of estimates of θ s.t.

$\mathrm{var}(\hat{\theta}_n) = O(\frac{1}{n})$, then this sequence of estimates is $\sqrt{n}$ - consistent.

Proof : By the definition of being bounded in probability and Chebyshev's inequality ,

$$P[\ |\hat{\theta}_n - \theta|\sqrt{n} \leq \varepsilon\ ] \geq 1 - \frac{\mathrm{var}(\hat{\theta}_n).n}{\varepsilon^2} ,$$

$$\geq 1 - \frac{k}{\varepsilon^2} - O(\frac{1}{n}) .$$

Theorem 2.4.2 : Let $\{\hat{\theta}_{1n}\}$ and $\{\hat{\theta}_{2n}\}$ be the sequences of maximum likelihood estimates and method of moments respectively, then $\{\hat{\theta}_{1n}\}$ and $\{\hat{\theta}_{2n}\}$ are $\sqrt{n}$ - consistent estimates.

Proof : If we could show that $\mathrm{var}(\hat{\theta}_{1n})$ and $\mathrm{var}(\hat{\theta}_{2n})$ are of $O(\frac{1}{n})$, then by the theorem 2.4.1 , these sequences of estimates are $\sqrt{n}$ - consistent. Following the methods in

Cramer(1946), Bickel and Docksum(1977), we have the following results.

i) the asymptotic covariance matrix of the maximum likelihood estimates is given by

$$\mathrm{var}(\hat{\theta}_{1n}) = \frac{1}{n}\mathcal{I}^{-1} \quad , \qquad (2.4.2)$$

where $\mathcal{I}$ is the inverse of the Fisher's information matrix obtained under the alternative hypothesis.

ii) Let $\hat{\theta}_{2n}$ be a function of first k raw(or central) moments $m_1,\ldots,m_k$ of the distribution f corresponding to the population moments $\mu_1,\ldots,\mu_k$.

Let $\hat{\theta}_{2i,n} = H(m_1,m_2,\ldots,m_k)$ then

$$\mathrm{var}(\hat{\theta}_{2i,n}) = \left[\frac{\partial H(\mu)}{\partial \mu}\right]' \Sigma^{(m)} \left[\frac{\partial H(\mu)}{\partial \mu}\right] , \qquad (2.4.3)$$

where $\mu = (\mu_1,\mu_2,\ldots,\mu_k)$,

$H(\mu) = H(\mu_1,\mu_2,\ldots,\mu_k)$ and

$$\Sigma^{(m)}_{ij} = \mathrm{Cov}(m_i,m_j) = O(\frac{1}{n}) \ .$$

Combining the results above we have ,

$$\mathrm{var}(\hat{\theta}_{2n}) = O(\frac{1}{n}) \ .$$

2.5 C(α) tests for testing of equality of parameters in presence of nuisance parameters

Consider k populations ,each characterised by a density function $f(\mu_i,\sigma_i^2)$, $i = 1,..,k$, μ_i and σ_i^2 are real numbers. Suppose n_i observations are taken from ith population and we wish to test the hypothesis of the equlity of μ_i's in presence of σ_i^2's.

Let $x_{ij} \sim f(\mu_i,\sigma_i^2)$, $\quad i = 1,..,k,\ j = 1,..,n_i$.

The competing hypotheses are

H_o: $\mu_1 = \mu_2 = ... = \mu_k$ and

H_A: at least one inequality

among μ_i's , (2.5.1)

σ_i^2's are unknown both under the null and the alternative hypotheses .

Reparametrise μ_i's as

$$\mu_i' = \mu + \phi_i \quad i = 1,..,k ,$$

with $\phi_k = 0$.

The new set of parameters is $(\phi_1,...,\phi_{k-1},\mu,\sigma_1^2,...,\sigma_k^2)$.

The ϕ_i's are related to the μ_i's by the transformation

$$\begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{k-1} \\ \mu \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \cdots & 0 & -1 \\ & & \vdots & & \\ 0 & 0 & & 1 & -1 \\ 0 & 0 & & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \\ \mu_{k-1} \\ \mu_k \end{bmatrix} ,$$

$$= A\underline{\mu} . \qquad (2.5.2)$$

The transformation A is 1-1 and onto , thus the testing of equality of μ_i's is equivalent to the following testing problem ,

H_o: $\phi_1 = \phi_2 = \ldots = \phi_{k-1} = 0$ and

H_A: at least one non-zero ϕ_i , (2.5.3)

where $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)$ is unknown nuisance parameter both under the null and the alternative hypotheses.

Let $\ell = \sum_i \sum_j \ln f(x_{ij}, \mu_i, \sigma_i)$,

$$= \sum_i \sum_j \ln f(x_{ij}, \mu + \phi_i, \sigma_i^2) \ . \qquad (2.5.4)$$

In the notation of section 2,

$\xi_1' = (\phi_1, \phi_2, \ldots, \phi_{k-1}) = \Phi'$,

$\xi_{10}' = (0, 0, \ldots, 0)$

and

$\xi_2' = (\mu, \sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)$

$= (\theta_1, \theta_2, \ldots, \theta_{k+1}) = \theta'$.

To derive the C(α) statistic, the scores and the mixed partial derivatives need to be evaluated under null hypothesis. Now,

i) Calculate $\frac{\partial \ell}{\partial \phi_i}$, $i = 1, .., k-1$, $\frac{\partial \ell}{\partial \mu}$,

$$\frac{\partial \ell}{\partial \sigma_i^2} \ , \ i = 1, .., k \ .$$

Define

$$\psi_i(\theta) = \frac{\partial \ell}{\partial \phi_i}\Big|_{\Phi = 0} ,$$

$$\gamma_1(\theta) = \frac{\partial \ell}{\partial \mu}\Big|_{\Phi = 0} \quad \text{and}$$

$$\gamma_{1+i}(\theta) = \frac{\partial \ell}{\partial \sigma_i^2}\Big|_{\Phi = 0} . \tag{2.5.5}$$

Also, $\frac{\partial \ell}{\partial \xi_1}\Big|_{\xi_1 = \xi_{10}} = (\psi_1, \psi_2, \ldots, \psi_{k-1})'$ and

$$\frac{\partial \ell}{\partial \xi_2}\Big|_{\xi_1 = \xi_{10}} = (\gamma_1, \gamma_2, \ldots, \gamma_{k+1}) .$$

ii)Calculate

$$\frac{\partial^2 \ell}{\partial \phi_i \partial \phi_j} = A^*_{ij} , \quad \frac{\partial^2 \ell}{\partial \phi_i \partial \mu} = B^*_{i1} , \quad \frac{\partial^2 \ell}{\partial \phi_i \partial \sigma_j^2} = B^*_{i,1+j} ,$$

$$\frac{\partial^2 \ell}{\partial \mu^2} = D^*_{11} , \quad \frac{\partial^2 \ell}{\partial \mu \partial \sigma_j^2} = D^*_{1,1+j} ,$$

$$\frac{\partial^2 \ell}{\partial \sigma_i^2 \partial \sigma_j^2} = D^*_{1+i,1+j} .$$

Define

$$A_{ij} = - E\left(A^*_{ij}\Big|_{\Phi = 0}\right) ,$$

$$B_{ij} = - E\left(B^*_{ij}\Big|_{\Phi = 0}\right)$$

and

$$D_{ij} = - E\left(D^*_{ij}\Big|_{\Phi = 0}\right) .$$

Again, $A = \Sigma_{11}$, $B = \Sigma_{12}$ and $D = \Sigma_{22}$ as in (2.3.4).

The vector of scores $(\psi_1, \psi_2, \ldots, \psi_{k-1}, \gamma_1, \gamma_2, \ldots, \gamma_{k+1})$ follows a multivariate normal distribution with expectation 0 and dspe. ion matrix

$$V = \begin{bmatrix} A & B \\ B' & D \end{bmatrix} ,$$

(Wald(1943), Rao(1948), etc). In this case dimensions of A, B and D are (k-1)x(k-1), (k-1)x(k+1) and (k+1)x(k+1), respectively.

Define $S = \Psi - BD^{-1}\Gamma$, then from (2.3.5) the C(α) statistic is

$$\chi^2_C = S'(A - BD^{-1}B')^{-1}S \sim \chi^2(k-1) \qquad (2.5.6)$$

where

$\Psi = (\psi_1, \psi_2, \ldots, \psi_{k-1})$ and

$\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_{k+1})$.

This method generalises the method described in Neyman(1959), Tarone(1985) etc. This does not impose more restrictions than requiring the expectations of mixed partial derivatives of log-likelihood to be finite. As this statistic still involves the nuisance parameters μ, σ_i^2's, they have to be replaced by their $\sqrt{n}$ - consistent estimates. Maximum likelihood estimators and method of moments estimators of μ and σ_i^2 are $\sqrt{n}$ - consistent, as shown in Thm. 2.4.2. By substituting the mle's of the

nuisance parameters in (2.5.6), χ^2_C reduces to a simpler form given by

$$\chi^2_{Cm} = \Psi'(A - BD^{-1}B')^{-1}\Psi \qquad (2.5.7)$$

which is asymptotically distributed as $\chi^2(k-1)$. However, if method of moments estimators are used the $C(\alpha)$ statistic is given by χ^2_C.

The block matrices A,B,D may have certain patterns depending upon the nature of the distribution. Frequently, the covariances between ψ_i's and $\gamma_2,\ldots,\gamma_{k+1}$ are zero, thereby producing a simple pattern in the information matrix V. Often it is found that $D_{ij} = 0$ for $2 \leq i \neq j \leq k+1$; as σ_i^2 and σ_j^2 are not related. A is a digonal matrix as ϕ_i and ϕ_j are not related. Three situations that arise commonly concerning σ_i^2 's are

i) When σ_i^2 's are known, in which case B is a column vector, D is a scalar and A is a diagonal matrix

ii) When $\sigma_i^2 = \sigma^2$, $i = 1,..,k$; σ^2 is unknown. In this case, B is a (k-1)x2 matrix,D is 2x2 matrx and A is again (k-1)x(k-1) matrix. Tarone's result(1985) can be obtained from (2.5.6) when D is a scalar d in which case B is a (k-1)x1 vector b, Γ is again a scalar variable γ, thus the statistic can be written as

$$\chi^2_C = (\Psi - \frac{1}{d}b\gamma)'(A - \frac{1}{d}bb')^{-1}(\Psi - \frac{1}{d}b\gamma)$$

iii) When σ_i^2 's are different, in this case B is (k-1)x(k+1); D is (k+1)x(k+1) matrix and A is again (k-1)x(k-1) matrix.

2.6 Detection of an outlier

Several explanations regarding the presence of outliers have been proposed (Barnett and Lewis(1978)).In this section, presence of one outlier is modelled as one observation coming from a distrbution having the same form of distribution but different parameters. This can also be formulated as a homogeneity testing problem with a restricted alternative hypothesis.

Consider y_i coming from a population with density $f(y,\mu_i)$. The competing hypotheses are

$$H_o: \mu_1 = \mu_2 = \ldots = \mu_k$$

and

$$H_A: \text{exactly one inequality among } \mu_i\text{'s} \qquad (2.6.1)$$

(one particular situation could be

$$\mu_1 = \mu_2 = \ldots = \mu_{k-1} = \mu > \mu_k \ )$$

This model will detect an upper outlier when the alternative has $\mu < \mu_k$ and it will detect a lower outlier when the alternative has $\mu > \mu_k$.

Reparametrise

$$\mu_i = \mu, \qquad i = 1,..,k-1$$

and

$$\mu_k = c\mu \ . \qquad (2.6.2)$$

Thus the competing hypotheses may be written as

$H_o: c = 1$ and

$H_A: c > 1$ (for an upper outlier) (2.6.3)

(or $c < 1$ for a lower outlier).

As this transformation is 1-1 and onto, the above two sets of hypotheses are equivalent.

Let $\ell = \sum_{i=1}^{k} \ln f(y_i, \mu_i) = \sum_{i=1}^{k} \ell_i$. Then

$$\frac{\partial \ell}{\partial c} = \frac{\partial \ell_k}{\partial c}, \quad \frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{k} \frac{\partial \ell_i}{\partial \mu} \quad , \tag{2.6.4}$$

$$\frac{\partial^2 \ell}{\partial \mu^2} = \sum_{i=1}^{k} \frac{\partial^2 \ell_i}{\partial \mu^2} = B^*,$$

$$\frac{\partial^2 \ell}{\partial \mu \partial c} = \frac{\partial^2 \ell_k}{\partial \mu \partial c} = A^* \text{ and} \tag{2.6.5}$$

$$\frac{\partial^2 \ell}{\partial c^2} = \frac{\partial \ell_k}{\partial c^2} = D^*.$$

Let $\psi(\mu) = \left.\frac{\partial \ell}{\partial c}\right|_{c=1}$

and $\gamma(\mu) = \left.\frac{\partial \ell}{\partial \mu}\right|_{c=1}$. (2.6.6)

also, $a = -E(A^* | c=1)$,

$b = -E(B^* | c=1)$,

$d = -E(D^* | c=1)$. (2.6.7)

Asymptotically (ψ, γ) follows a bivariate normal distribution with expectation 0 and dispersion matrix

$V = \begin{bmatrix} d & a \\ a & b \end{bmatrix}$. The conditional distribution of ψ given γ

is a univariate normal with mean $\beta\gamma$ and

variance $(d - \frac{a^2}{b})$, where $\beta = \frac{a}{b}$.

Thus the test-statistic for testing no outliers against a one-sided alternative is

$$z = \frac{\sqrt{b}\ (\psi - \beta\gamma)}{\sqrt{bd - a^2}}\ , \qquad (2.6.8)$$

which is asymptotically distributed as N(0,1). Note that the test is one-sided, so at level 100α %, H_o is rejected in favour of an upper outlier if $z > z_{\alpha}$, and similarly H_o is rejected in favour of a lower outlier if $z < z_{\alpha}$. Two-sided tests of the outlier are not performed as it lacks proper physical interpretation. By the inspection of the data we could see whether an upper outlier or lower outlier is suspected.

This model can also be interpreted as the 'shock model' in a control chart. As T still involves nuisance parameter μ, we need to substitute a $\sqrt{n}$-consistent estimate for μ in the above expression. This formulation can also be used in situations where (k-1) treament effects are found the same, and it is desirable to check if the treatments are in any way different from the control group .

2.7 Testing of a set of linear constraints

Frequently , instead of testing equality of certain parameters of several distributions, it is required to test whether m (< k) combinations of these parameters have certain specified values (see Bühler and Puri(1966), Rao(1948)).

Let $y_{ij} \sim f(y,\mu_i)$, and the log-likelihood of the data

$$\ell = \sum_i \sum_j \ln f(y_{ij},\mu_i) \ ,$$

$$i = 1,..,k,\ j = 1,..,n_i, \qquad (2.7.1)$$

where f is a valid probability distribution s.t. lnf is a Cramer function. Let ℓ be the log-likelihood of the data set. Define $\mu' = (\mu_1,\mu_2,\ldots,\mu_k)$.

Let the m equality constraints be

$$a_1'\mu = A_{11}\mu_1 + A_{12}\mu_2 + \ldots + A_{1k}\mu_k = \theta_1 \ ,$$

$$a_2'\mu = A_{21}\mu_1 + A_{22}\mu_2 + \ldots + A_{2k}\mu_k = \theta_2 \ ,$$

$$\ldots. \qquad (2.7.2)$$

$$\ldots.$$

$$a_m'\mu = A_{m1}\mu_1 + A_{m2}\mu_2 + \ldots + A_{mk}\mu_k = \theta_m \ ,$$

alternatively written as $A\mu = \theta_0$(a vector of known constants), and A is an mxp full rank matrix. Augment the matrix A by adding rows from the standard basis of $\mathcal{R}^k$ such that the resulting matrix is a non-singular kxk matrix. The rows of the matrix are defined as

$\xi'_{1i} = a'_i\mu$, $i = 1,..,m,$

$\xi'_{2j} = e'_j\mu$, $j = 1,..,k-m,$ where (2.7.3)

e_j is a member of the the standard basis of $\mathcal{R}^k$ with 1 at the jth place and zero in other places.

The new parametrisation is

$$\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} \\ 0 & I \end{bmatrix}\mu = B\mu \qquad (2.7.4)$$

Where ξ_1 is an mx1 vector and ξ_2 is a (k-m)x1 vector. By rearranging the rows of matrix A we arrive at the above matrix B.

The usual testing problem is

$$H_o: A\mu = \theta_o$$

and (2.7.5)

$$H_A: A\mu = \theta_o .$$

As B is a non-singular matrix , which makes the transformation from μ to ξ as 1-1 and onto. Substituting $B^{-1}\xi$ for μ in ℓ will provide an expression of ℓ in terms of ξ. Thus the new testing problem can be formulated as

$$H_o: \xi_1 = \theta_o$$

and (2.7.6)

$$H_A: \xi_1 \neq \theta_o .$$

The parameter μ and ξ are related as

$$\mu = \begin{bmatrix} B_{11}^{-1}\xi_1 \\ \xi_2 - B_{11}^{-1}B_{12}\xi_1 \end{bmatrix}. \qquad (2.7.7)$$

Define the following derivatives of ℓ after substituting $\mu = B^{-1}\xi$ in (2.7.1),

$$\psi_i(\xi_2) = \frac{\partial \ell}{\partial \xi_{1i}}\Big|_{\xi_1 = \theta_0}, \qquad i = 1,..,m,$$

$$\gamma_i(\xi_2) = \frac{\partial \ell}{\partial \xi_{2i}}\Big|_{\xi_1 = \theta_0}. \qquad i = 1,..,k-m. \qquad (2.7.8)$$

The information matrix V consists of the following block matrices D,B,A as

$$V = \begin{bmatrix} D & A \\ A' & B \end{bmatrix}, \text{ where} \qquad (2.7.9)$$

$$D_{ij} = -E\left[\frac{\partial^2 \ell}{\partial \xi_{1i}\partial \xi_{1j}}\Big|_{\xi_1 = \theta_0}\right], \qquad 1 \le i,j \le m,$$

$$A_{ij} = -E\left[\frac{\partial^2 \ell}{\partial \xi_{1i}\partial \xi_{2j}}\Big|_{\xi_1 = \theta_0}\right],$$

$$1 \le i \le m;\ 1 \le j \le k-m,$$

and

$$B_{ij} = -E\left[\frac{\partial^2 \ell}{\partial \xi_{2i}\partial \xi_{2j}}\Big|_{\xi_1 = \theta_0}\right],$$

$$1 \le i,j \le k-m.$$

Now define adjusted scores

$$S_i = \psi_i - \sum_{j=1}^{k-m} \beta_{ij}\gamma_j \ , \qquad i = 1,\ldots,m \ , \qquad (2.7.10)$$

ψ_i and γ_i stand for $\psi_i(\xi_2)$ and $\gamma_i(\xi_2)$.

The vector $(\psi_1,\psi_2,\ldots,\psi_m,\gamma_1,\ldots,\gamma_{k-m})$ follows asymptotically multivariate normal distribution with expectation 0 and the covariance matrix V. Thus the test-statistic for this hypothesis is obtained following the same arguments as before

$$\chi^2_A = S'(D - AB^{-1}A')^{-1}S \sim \chi^2(m) \qquad (2.7.11)$$

As χ^2_A involves the nuisance parameter ξ_2 which is unknown, thus, in order to compute the test-statistic from data we replace ξ_2 by its $\sqrt{n}$ - consistent estimate. Further simplification of this statistic is possible depending upon the structure of the matrix A.

2.8 WALD'S TEST :

Wald(1943) studied in detail the asymptotic test of composite hypotheses by considering the distribution of the maximum likelihood estimates. In the setting of the earlier sections , let ℓ be the log-likelihood involving the vector of parameters (ξ_1,ξ_2) s.t. ξ_1 is sx1 vector and ξ_2 is (p-s)x1 vector. Instead of defining a density for each individual x_{ij}, the joint density of all the observations are taken together in order to cover different hypotheses concerning the relation among the parameters apart from the usual testing of homogeneity of some parameter.The usual competing hypotheses are

H_0: $\xi_1 = \xi_{10}$ and

H_A: $\xi_1 \neq \xi_{10}$, (2.8.1)

where ξ_2 is unspecified both under H_0 and H_A. Let $\hat{\xi}_1$ and $\hat{\xi}_2$ be the maximum likelihood estimates of ξ_1 and ξ_2 under the alternative hypothesis. It is proved that

$$(\hat{\xi}_1, \hat{\xi}_2) \sim MN((\xi_1, \xi_2), \mathcal{I}^{-1}) , \quad (2.8.2)$$

where $\mathcal{I}^{-1}$ is the inverse of the information matrix under the alternative hypothesis. The Wald's statistic is

$$\chi^2_W = \sum_i \sum_j \mathcal{I}^{ij} (\hat{\xi}_{1i} - \xi_{10i})(\hat{\xi}_{1j} - \xi_{10j}) ,$$

$$= (\hat{\xi}_1 - \xi_{10})' \mathcal{I}^{-1}_{11.2} (\hat{\xi}_1 - \xi_{10}) , \quad (2.8.3)$$

where $\mathcal{I}_{11.2}$ involves the estimate $(\hat{\xi}_1, \hat{\xi}_2)$.

The derivation of this test-statistic parallels the derivation of the $C(\alpha)$ test. Using (2.8.2)

$$\hat{\xi}_1 | \hat{\xi}_2 \sim MN(\xi_1 - B(\hat{\xi}_2 - \xi_2), \mathcal{I}_{11.2}) , \quad (2.8.4)$$

where

$B = \mathcal{I}_{12} \mathcal{I}^{-1}_{22}$, matrix of partial regression coefficients .

Define $S = (\hat{\xi}_1 - \xi_1) - B(\hat{\xi}_2 - \xi_2)$ then by (2.8.4), the statistic given by

$$\chi^2_W = S' \mathcal{I}^{-1}_{11.2} S \quad (2.8.5)$$

is distributed as χ^2 with s d.f.

If ξ_2 is replaced by $\hat{\xi}_2$ and ξ_1 by ξ_{10} then the statistic in (2.8.5) reduces to the statistic in (2.8.3). Let χ^2_L and χ^2_L be the likelihood ratio test and the C(α) test for the setting in (2.8.1). Now,

$$\chi^2_L = -2(\ell(\hat{\xi}_1, \hat{\xi}_{12}) - \ell(\xi_{10}, \hat{\xi}_{02})), \qquad (2.8.6)$$

where

$\hat{\xi}_{02}$ = maximum likelihood estimate of ξ_2 under H_0 ,

$\hat{\xi}_{12}$ = maximum likelihood estimate of ξ_2 under H_A

and

$\chi^2_{C\alpha}$ derived using the method described earlier. .

Kendall & Stuart(1979), show that asymptotically

$$\chi^2_L = (\hat{\xi}_1 - \xi_{10})' \mathcal{I}^{-1}_{11.2} (\hat{\xi}_1 - \xi_{10}) , \qquad (2.8.7)$$

which is χ^2_W in (2.8.5). This shows that χ^2_L and χ^2_W are asymptotically equivalent. Chant(1974) and Moran(1970) show that $\chi^2_{C\alpha}$ and χ^2_W are asymptotically equivalent. Combining these facts, we get the asymptotic equivalence of χ^2_L, χ^2_W and $\chi^2_{C\alpha}$.

The use of the C(α) test should be preferred to the likelihood ratio test and Wald's test for the following reasons :

i) Covariance matrix of the scores is exactly equal to the Fisher's information matrix whereas the covariance matrix of maximum likelihood estimates is asymptotically equal to the inverse of the information matrix.

ii) Wald's statistic requires calculation of maximum likelihood estimates of all the parameters of the model whereas the $C(\alpha)$ test requires estimation of the nuisance parameters only.

iii) Method of moments estimates can be used in the $C(\alpha)$ test whereas by definition we can not use these estimates in the Wald's statistic.

iv) Likelihood ratio test statistic does not impose many conditions like regularity of density function, range being independent of the parameter etc. so far as the construction goes. However the proof of the convergence of the test statistic to a χ^2 random variable is acheived by making the same assumptions as needed to derive the Wald's test as discussed in Kendall & Stuart(1979). However, the $C(\alpha)$ test may not be the best test for all the problems of testing of the composite hypotheses. But, asymptotically, it does provide a critical region even in complicated situations.

CHAPTER 3

TESTING OF EQUALITY OF GROUP MEANS FOR POISSON COUNT DATA AND DETECTION OF NEGATIVE BINOMIAL VARIATION

3.1 Introduction

In several practical situations ,it is required to compare M treatments or M groups or M different levels of dosage etc. A typical data set could be presented as

groups	observations
1	$x_{11}, x_{12}, \ldots, x_{1n_1}$
2	$x_{21}, x_{22}, \ldots, x_{2n_2}$
.	.....
.	.....
.	.....
M	$x_{M1}, x_{M2}, \ldots, x_{Mn_M}$

Thus x_{ij} = jth observation in the ith group.

The usual analysis of variance model is

$$x_{ij} = \mu + \tau_i + e_{ij} ,$$

where

$e_{ij} \sim$ IID $N(0, \sigma^2)$,

such that $\sum_i \tau_i = 0$ (3.1.1)

and $\mu , \tau_i , e_{ij} \in \mathbb{R}$.

We are interested in testing

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_M$$

against

H_A : at least one inequality among τ_i's . (3.1.2)

However , this formulation fails when the observations are counts or discrete data.

As the parameters μ and τ_i's are real, it will be very difficult to obtain a distribution for e_{ij} which will make the observations x_{ij} discrete.

Light and Margolin (1971) discuss analysis of variance for categorical data as observations falling in a two-dimensional contingency table with one margin fixed, which is very limited in applications.

This problem can be reformulated in terms of

$$E(y_{ij}) = \mu + \tau_i = \mu_i,$$

hence , (3.1.2) is equivalent to testing

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_M ,$$

against

H_A : at least one inequality in μ_i's

in the presence of nuisance parameters.

(3.1.3)

Hence , the usual testing of equality of treatment effects is equivalent to testing the equality of group means .This formulation incorporates both continuous and discrete distributions . Generally speaking , this is a problem of testing homogeneity of certain parameter in one-way classification in the presence of nuisance parameter. Normal theory ANOVA problem is testing of homogeneity of group means when the observations are independently

distributed as normal variates with common unknown variance (the nuisance parameter). In this chapter, we develop tests for equality of means under Poisson distribution and detection of extra-poisson variation in one-way layout of count data.

3.2. Testing Of Homogeneity Under Poisson Variation

In the absence of any information regarding the underlying distribution of count data it is customary to assume that the observations are coming from Poisson distribution. The testing of homogeneity of Poisson parameters has been discussed in Pothoff and Whittinghall (1966) and Moran (1973). These techniques are derived for detecting the extra-poisson variation in a single sample of data. Plackett(1981) and Kendall & Stuart(1979) have mentioned one-way layout of count data very briefly. In this section, the usual test statistics are presented in formal set-up employing the well-known methods (see Bishop, Fienberg & Holland, 1975).

Consider count data set

$$\{ x_{ij} ; i = 1,\ldots, M; j = 1,\ldots, n_i \},$$

s.t. $x_{ij} \sim Poi(\lambda_i)$. (3.2.1)

The competing hypotheses for equality of means are

$$H_0 : \lambda_1 = \lambda_2 = \ldots = \lambda_M$$

and

H_A : at least one inequality

among λ_i's. (3.2.2)

Three types of statistics for testing H_o against H_A are presented.

i) Likelihood ratio test:

The log-likelihood ℓ_1 when all the λ_i's are possibly unequal is given by

$$\ell_1 = -\sum_i n_i\lambda_i + \sum_i x_{i+} \ln\lambda_i . \quad (3.2.3)$$

The log-likelihood ℓ_o under H_o is given by

$$\ell_o = -n.\lambda + x_{++}\ln\lambda, \quad (3.2.4)$$

where

$$x_{++} = \sum_i x_{i+} = \sum_i\sum_j x_{ij}$$

and

$$n = \sum_i n_i .$$

The usual LR test is $2(\ell_1 - \ell_o)$,evaluated by substituting the values of maximum likelihood estimates in place of the parameters .

$$\chi_1^2 = -2(\hat{\ell}_o - \hat{\ell}_1)$$

$$= -2(\sum n_i\bar{x}_i \ln(\bar{x}_i / \bar{x})) \quad (3.2.5)$$

ii) Conditional tests : Due to the reproductive property of Poisson distribution and (3.2.1),

$$x_{i+} \sim \text{Poi}(n_i\lambda_i)$$

and

$$x_{++} \sim \text{Poi}(\sum_i n_i\lambda_i).$$

Conditional on x_{++} , the x_{i+}'s are distributed as Mult(x_{++};p_1, p_2, ... ,p_M) where $p_i = n_i\lambda_i / (\sum n_i\lambda_i)$.

Under the null hypothesis, $p_i = n_i/n$. Hence the usual goodness-of-fit statistic $\sum_i (O_i - E_i)^2/E_i$, yields

$$\chi_2^2 = \sum_i n_i(\bar{x}_i - \bar{\bar{x}})^2 / \bar{\bar{x}} . \qquad (3.2.6)$$

Again, conditional on x_{++}, the x_{ij}'s are jointly distributed as Mult(x_{++}; p_{ij}),

where

$$p_{ij} = \lambda_i / \Sigma n_i\lambda_i , \qquad i = 1,..,M,\ j = 1,..,n_i .$$

Under the null hypotheses, $p_{ij} = 1/n$. The goodness-of-fit statistic in this case is

$$\chi_3^2 = \sum_i\sum_j (x_{ij} - \bar{\bar{x}})^2 / \bar{\bar{x}} \qquad (3.2.7)$$

$$= \Sigma\Sigma(x_{ij} - \bar{x}_i)^2 / \bar{\bar{x}} + \chi_2^2$$

$$= \sum_i (\bar{x}_i / \bar{\bar{x}}) \sum_j (x_{ij} - \bar{x}_i)^2 / \bar{x}_i + \chi_2^2 . \qquad (3.2.8)$$

The statistics χ_1^2 and χ_2^2 are asymptotically distributed as χ^2 with d.f.(M-1) whereas χ_3^2 is distributed as χ^2 with d.f. (n-1). The statistic χ_3^2 is an extended index of dispersion test. The statistic χ_2^2 is similar to the F-statistic in normal theory ANOVA in which, it is used to test homogeneity of means. However, χ_3^2 is more informative than the other two statistics as it can be partitioned and explained in two ways

a) The first part is the statistic T_c in Collings & Margolin (1985), for detecting a departure from the assumption of Poisson distribution, and the second part is

χ^2_2, which is used to test homogeneity of means..

b) referring to (3.2.8) , the first part is the weighted sum of the indices of dispersion of the M groups , and the second is χ^2_2 .

A test based on χ^2_3 should be used first . If this is found significant, one should then check if one or both components are significant. If both are significant then data gives evidence of heterogeneity of means and extra-poisson variation.

iii) C(α)Tests :

For constructing a C(α) test for this situation , reparametrise λ_i's as discussed in Chapter 2, as

$$\lambda_i = \lambda + \phi_i, \qquad i = 1,..,M,$$

$$\text{s.t. } \phi_M = 0. \tag{3.2.9}$$

Then the hypotheses reduce to

$$H_0 : \phi_1 = \phi_2 = \ldots = \phi_M$$

and

$$H_A : \text{at least one inequality in } \phi_i\text{'s} \tag{3.2.10}$$

Consider the log-likelihood ℓ under unequal ϕ_i's ,

$$\ell = -\Sigma n_i(\lambda + \phi_i) + \Sigma x_{i+} \ln(\lambda + \phi_i). \tag{3.2.11}$$

Differentiate ℓ with respect to ϕ_i's and λ and substitute 0 for ϕ_i's in the resulting expressions :

$$\left.\frac{\partial \ell}{\partial \phi_i}\right|_{\Phi = 0} = -n_i + x_{i+} ,$$

$$= \psi_i(\lambda) \qquad i = 1,\ldots,M-1,$$

and

$$\frac{\partial \ell}{\partial \lambda}\Big|_{\Phi = 0} = -n + x_{++} = \gamma(\lambda). \qquad (3.2.12)$$

Again, differentiating ℓ twice with respect to ϕ_i's and λ, setting ϕ_i's equal to 0 in the mixed partial derivatives then taking the expectation of the resulting expression under the null hypotheses :

$$E(-\partial^2 \ell / \partial\phi_i \partial\phi_j |_{\Phi=0}) = n_i/\lambda , \qquad 1 \le i = j \le M-1,$$

$$= 0 , \qquad \text{otherwise,}$$

$$E(-\partial^2 \ell / \partial\phi_i \partial\lambda |_{\Phi=0}) = n_i/\lambda, \qquad i = 1,\ldots,M-1,$$

$$E(-\partial^2 \ell / \partial\lambda^2 |_{\Phi=0}) = n/\lambda. \qquad (3.2.13)$$

For large n, $(\psi_1,\psi_2,\ldots,\psi_{k-1},\gamma)' \sim N(0, V)$

where,

$$V = \begin{bmatrix} A & B \\ B' & D \end{bmatrix}, \qquad (3.2.14)$$

$$A = -\frac{1}{\lambda} \text{ Diag}(n_1,n_2,\ldots, n_M),$$

$$B = -\frac{1}{\lambda}(n_1,n_2, \ldots ,n_M)$$

and

$$D = n/\lambda .$$

Define $S_i = \psi_i - \beta_i\gamma$ (ψ_i's and γ are functions of λ). The marginal distribution of

$Y = (S_1, S_2, .., S_{M-1})' \sim N(0, A - BB'/D)$.

Thus the C(α) test is given by

$$\chi_4^2 = Y'(A - BB'/D)^{-1}Y \sim \chi^2(M-1). \qquad (3.2.15)$$

After some simplification ,

$$\chi_4^2 = \Sigma\, n_i(\bar{x}_i - \lambda)^2 / \lambda \ . \qquad (3.2.16)$$

Substituting $\bar{x}$ for λ in (3.2.16),

$$\chi_4^2 = \Sigma\, n_i(\bar{x}_i - \bar{x})^2 / \bar{x} \ . \qquad (3.2.17)$$

Thus the usual goodness-of-fit statistic for testing the homogeneity of Poisson parameters not only possesses the optimal properties of a conditional test , it is also locally asymptotically most powerful.

3.3 Detection of negative binomial variation

In many statistical analyses, negative binomial distribution is taken as an alternative to the Poisson distribution for a count data showing over-dispersion (see Margolin (1985), Macaughrn & Arnold (1976)). However, before analysing the data, assuming that observations follow negative binomial distribution, we must check if there is any evidence of negative binomial over-dispersion. The probability mass function of a random variable Y following a negative binomial distribution with mean m and dispersion parameter c,

denoted by $Y \sim NB(m,c)$, is

$$P(Y = y) = \frac{\Gamma(y + c^{-1})}{y!\ \Gamma(c^{-1})} \left(\frac{cm}{1+cm}\right)^{y} \left(\frac{1}{1+cm}\right)^{c^{-1}} ,$$

for $y = 0,1,\ldots.$; $0 < m < \infty$; $0 < c < \infty$. Here

$E(Y) = m$ and $var(Y) = m + cm^2$. Evidently , the $NB(m,c)$ distribution tends to Poisson distribution with parameter m in the limit when $c \longrightarrow 0$. Then to check if there is any evidence of departure from Poisson distribution, in favour of negative binomial distribution is to test

$$H_o : c = 0$$

against

$$H_A : c > 0 \qquad (3.3.1)$$

For this we develop several $C(\alpha)$ tests and a likelihood ratio test and compare them in terms of size and power by simulation.

i) $C(\alpha)$ tests :

Let $y_{ij} \sim NB(m_i, c)$, $\quad i = 1,\ldots,k$, $j = 1,\ldots,n_i$.

The log-likelihood of all the observations under H_A is

$$\ell = \sum_i \ell_i \ . \qquad (3.3.2)$$

Similarly , all the first and second derivatives can be calculated as

$$\frac{\partial \ell}{\partial c} = \sum_i \partial \ell_i / \partial c \ ,$$

$$\partial^2 \ell / \partial c^2 = \sum_i \partial^2 \ell_i / \partial c^2 ,$$

$$\partial \ell / \partial m_i = \sum \partial \ell_i / \partial m_i = \partial \ell_i / \partial m_i ,$$

$$\partial^2 \ell / \partial m_i \partial c = \partial^2 \ell_i / \partial m_i \partial c ,$$

and

$$\partial^2 \ell / \partial m_i^2 = \partial^2 \ell_i / \partial m_i^2 .$$

(3.3.3)

Hence calculating the derivatives in (3.3.3) requires differentiating the log-likelihood of ith population with respect to its parameters. This calculation can further be simplified by considering a set of observations $y_1, y_2, \ldots, y_n$ from NB(m,c) , calculate first and second derivatives of log-likelihood of this data and then place the index i suitably. This way, the calculations of (3.3.3) are greatly simplified for better presentation.

The log-likelihood for $y_i \sim$ NB(m,c) for $i = 1, \ldots, n$ is given by

$$\ell = \sum_{1}^{n} \sum_{0}^{y_i - 1} \ln(j + c^{-1}) - \frac{n}{c} \ln(1 + cm) + \sum y_i (\ln c + \ln m - \ln(1 + c.m)). \qquad (3.3.4)$$

The derivatives are :

$$\frac{\partial \ell}{\partial c} = \frac{n \ln(1 + cm)}{c^2} - \frac{1}{c} \sum_i \sum_j \frac{1}{(1 + cj)} + \frac{\Sigma y_i - nm}{c(1 + cm)} , \qquad (3.3.5)$$

$$\frac{\partial \ell}{\partial m} = \frac{n(\bar{y} - m)}{m(1 + cm)} , \qquad (3.3.6)$$

$$\frac{\partial^2 \ell}{\partial c^2} = \frac{n\ m}{c^2(1+cm)} + \frac{2}{c^2} \sum\sum \frac{1}{(1 + cj)}$$

$$- \frac{n}{c^2} \ln(1 + cm) - \frac{1}{c^2} \sum\sum \frac{1}{(1 + cj)^2}$$

$$+ \frac{n(\bar{y} - m)(1 + 2cm)}{c^2(1+cm)^2} , \qquad (3.3.7)$$

$$\frac{\partial^2 \ell}{\partial m\ \partial c} = - \frac{n(\bar{y} - m)}{(1+cm)^2} ,$$

$$\frac{\partial^2 \ell}{\partial m^2} = - \frac{n}{m(1+cm)} + \frac{n(1+2cm)(\bar{y} - m)}{m^2(1+cm)^2} , \qquad (3.3.8)$$

where m is the nuisance parameter .

First we calculate $\partial \ell / \partial c$ at c = 0. This expression is in indeterminate form, thus applying L'Hospital rule

$$T = \lim_{c \to 0} \frac{\partial \ell}{\partial c}$$

$$= \sum y_i(y_i - 1)/2 - \sum m y_i + nm^2/2. \qquad (3.3.9)$$

Substituting $\bar{y}$ for m in (3.3.9) , we get

$$T = \frac{n}{2}(S^2 - \bar{y}) \qquad (3.3.10)$$

where

$$S^2 = \frac{1}{n} \sum_i (y_i - \bar{y})^2 .$$

Similarly, from Collings(1981) and Fisher(1941) , we find

$$E\left[- \frac{\partial^2 \ell}{\partial c^2} \Big|_{c = 0} \right] = \frac{n\, m^2}{2} , \qquad (3.3.11)$$

$$E\left[- \frac{\partial^2 \ell}{\partial m\, \partial c} \Big|_{c = 0} \right] = 0$$

and

$$E\left[- \frac{\partial^2 \ell}{\partial m^2} \Big|_{c = 0} \right] = \frac{n}{m} .$$

Define

$$\psi_i = \frac{\partial \ell}{\partial m} \Big|_{c = 0}$$

and

$$\gamma_i = \frac{\partial \ell}{\partial c} \Big|_{c = 0} .$$

Asymptotically the joint distrbution of ψ_i and γ_i is bivariate normal with expectation 0 and covariance matrix V given by

$$V = \begin{bmatrix} \frac{n\, m^2}{2} & 0 \\ 0 & \frac{n}{m} \end{bmatrix} . \qquad (3.3.12)$$

As the covariance (V_{12}) in (3.3.12) is zero , the conditional distribution of ψ_i given γ_i is again normal with mean 0 and variance $n\, m^2/2$. Replacing m by its

root-n consistent estimator $\bar{y}$ and standardising T, we obtain the C(α) statistic for testing negative binomial overdispersion

$$\chi_1 = \frac{\sqrt{n/2}\ (S^2 - \bar{y}\ )}{\bar{y}} \ . \qquad (3.3.13)$$

This has also been derived by Zelterman and Chen(1988). The usual approximation of the distribution of the statistic by a χ^2 will not be valid as the alternative in our case is one-sided whereas the critical region determined by a χ^2 distribution is two-sided. For this reason, we have denoted the test-statistic by χ_1, and compare this statistic by the percentage points of a standard normal distribution. This statistic is a linear function of the usual index of dispersion test. While index-of-dispersion test detects extra-poisson variation, this statistic detects negative binomial variation.

Carefully using the results obtained earlier , we derive the C(α) test for detection of negative binomial overdispersion in one-way layout. Thus ,

$$T = \lim_{c \to 0} \frac{\partial \ell}{\partial c} = \sum_i \lim_{c \to 0} \frac{\partial \ell_i}{\partial c} \ .$$

$$= \frac{1}{2} \sum_i \sum_j \left((y_{ij} - m_i)^2 - y_{ij} \right)$$

$$- \sum_i m_i \left(\sum_j y_{ij} - n_i m_i \right) , \qquad (3.3.14)$$

$$\frac{\partial \ell}{\partial m_i} = \frac{n_i (\bar{y}_i - m_i)}{m_i (1 + c m_i)} ,$$

$$\frac{\partial^2 \ell}{\partial c \, \partial m_i} = - \frac{n_i (\bar{y}_i - m_i)}{(1 + c m_i)^2} ,$$

$$\frac{\partial^2 \ell}{\partial c^2} = \sum \sum \left\{ \frac{m_i}{c^2 (1+c m_i)^2} + \frac{2}{c^2} \sum_t \frac{1}{1+ ct} \right.$$

$$- \frac{\ln(1+c m_i)}{c^2} \qquad - \frac{1}{c^2} \sum_t \frac{1}{(1+ct)^2} \Bigg\}$$

$$+ \sum_i \frac{n_i (\bar{y}_i - m_i) (1 + 2 c m_i)}{c^2 (1 + c m_i)^2} , \qquad (3.3.15)$$

$$\frac{\partial^2 \ell}{\partial m_i^2} = \frac{n_i}{m_i (1 + c m_i)}$$

and

$$\frac{\partial^2 \ell}{\partial m_i \partial m_j} = 0 .$$

The dispersion matrix of $\left(\frac{\partial \ell}{\partial c} , \frac{\partial \ell}{\partial m_1} , \ldots , \frac{\partial \ell}{\partial m_k} \right)$ is given by

$$V = \begin{bmatrix} A & B \\ B' & D \end{bmatrix} , \qquad (3.3.16)$$

where A is a scalar, D is a kxk diagonal matrix and B is a kx1 vector as described below.

$$A = E\left(-\frac{\partial^2 \ell}{\partial c^2}\right) = \sum_i n_i m_i^2 / 2 \ , \qquad (3.3.17)$$

$$B_i = E\left[-\frac{\partial^2 \ell}{\partial c \partial m_i} \Big|_{c=0} \right] = 0, \qquad i = 1,\ldots,k,$$

$$D_{ij} = -E\left[\frac{\partial^2 \ell}{\partial m_i \partial m_j} \Big|_{c=0} \right], \qquad 1 \le i,j \le k$$

and

$$D_{ij} = 0 \qquad 1 \le i \ne j \le k$$
$$= n_i/m_i \quad \text{otherwise}$$

Hence the conditional distribution of T in (3.3.14) is approximately N(0,A). Substituting $\bar{y}_i$ for m_i in the expressions (3.3.14) and (3.3.17), the C(α) statistic is

$$T_1 = T/\sqrt{A} \ ,$$

$$= \frac{\sum n_i (S_i^2 - \bar{y}_i)}{\sqrt{2 \sum n_i \bar{y}_i^2}} \ . \qquad (3.3.18)$$

Once again, due to the one-sided alternative hypotheses, the statistic T_1 should be compared with standard normal percentage points. Collings & Margolin (1985) develop a

test criterion for (3.3.1) which is

$$T_c = \Sigma \Sigma (y_{ij} - \bar{y}_i)^2 / \bar{y} \ . \qquad (3.3.19)$$

Under H_0 , $T_c \sim \Sigma w_i \chi_i^2 (n_i - 1)$ where $w_i = \bar{y}_i / \bar{y}$. As we can see, a test based on T_1 is easier to perform than on T_c as the critical values of T_c is data dependent and not easily available.

A finer approximation of T_1 can be obtained by calculating the first and second moments of T_1 and defining

$$T_3 = \frac{T - E(T)}{sd(T)} \ ,$$

where

$$T = \frac{1}{2} \Sigma n_i (S_i^2 - \bar{y}_i) = \frac{1}{2} \Sigma n_i D_i \ .$$

Now, $E(T) = \frac{1}{2} \Sigma n_i E(D_i)$ and $var(T) = \frac{1}{4} \Sigma n_i^2 var(D_i)$. Under H_0 ,

$$E(D_i) = - \frac{m_i}{n_i}$$

and

$$var(D_i) = 2(n_i - 1) m_i^2 + m_i / n_i \ .$$

Hence the new statistic corrected for mean and variance is

$$T_3 = \frac{\Sigma n_i (D_i + m_i / n_i)}{\sqrt{2 \Sigma ((n_i - 1) m_i^2 + m_i / 2n_i)}} \ .$$

Replacing m_i's by their maximum likelihood estimates $\bar{y}_i$'s, we obtain

$$T_3 = \frac{\sum n_i (S_i^2 - \bar{y}_i + \bar{y}_i/n_i)}{\sqrt{2\Sigma((n_i-1)\bar{y}_i^2 + \bar{y}_i/2n_i)}}, \qquad (3.3.20)$$

which is also asymptotically N(0,1).

ii) Range justified tests : The testing problem concerning the detection of extra-poisson variation (negative binomial overdispersion) is one-sided and the value of the dispersion parameter is restricted to non-negative real numbers only.

Appealing to the non-negative restriction of the parameter, two more statistics are defined for detection in one-way layout as follows

$$T_2 = \frac{\sum_i n_i (S_i^2 - \bar{y}_i)_+}{\sqrt{\sum_i 2n_i \bar{y}_i^2}} \qquad (3.3.21)$$

$$T_4 = \frac{\sum_i (S_i^2 - \bar{y}_i + \bar{y}_i/n_i)_+}{\sqrt{2\Sigma_i((n_i-1)\bar{y}_i^2 + \bar{y}_i/n_i)}} \qquad (3.3.22)$$

where $A_+ = \max(A,0)$ where A is any real number.

iii) Likelihood Ratio test :

The usual LR test $2(\ell(c) - \ell(0))$ is distributed as χ^2 with 1 d.f. , where

$\ell(c)$ = likelihood of data under the assumption of negative binomial distribution evaluated at the maximum likelihood estimates of the parameters ,

$\ell(0)$ = likelihood of the data under the assumption of Poisson distribution evaluated at the maximum likelihood estimates of the parameters .

This statistic has the following complicated expression

$$\frac{1}{2}\chi^2_\sigma = \sum_i \sum_j \{ \sum_{t=0}^{y_{ij}-1} \ln(1+c_1 t) \} - y_{ij} \ln(1+c_1 \bar{y}_i) \}$$
$$+ \sum \frac{n_i \{ \ln(1+c_1 \bar{y}_i) - c_1 \bar{y}_i \}}{c_1} , \qquad (3.3.23)$$

where c_1 is the maximum likelihood estimate of the dispersion parameter c. Expanding χ^2_σ as Taylor series about $c_1 = 0$, we get

$$\chi^2_\sigma = c_1 (\sum n_i (S_i^2 - \bar{y}_i)) + O(c_1^2) \qquad (3.3.24)$$

χ^2_σ will reject the hypothesis of absence of negative binomial variation for large values of $\sum n_i (S_i^2 - \bar{y}_i)$.

Since the distribution of this statistic is intractable, we use the asymptotic distribution derived earlier. The value of c is obtained by solving the following equation iteratively

$$\sum \frac{n_i \ln(1+c\bar{y}_i)}{c^2} - \sum \sum \sum_{t=0}^{y_{ij}-1} \frac{1}{c(1+ct)} = 0 .$$

The statistic T_2 is superior to its competitors computationally as

a) T_C requires calculation of percentage points from a mixture of χ^2 with different d.f.s.These percentage points are not readily available.

b) χ^2_σ requires the estimate of the dispersion parameter obtained by solving the above complicated equation. From the Taylor series expansion , it is clear that T_C and χ^2_σ are to a large extent funnctionally related (at least locally).

From eqn.(3.3.23) we find that as $c \longrightarrow 0$, χ^2_σ converges to a degenerate r.v. taking value 0 with probability 1. Following Self and Liang(1987) and Paul et al.(1989) , another approximation to the χ^2_σ is a 50:50 mixture of $\chi^2(0)$ and $\chi^2(1)$. This simply means that χ^2_σ be compared with one-half of the percentage point ofχ^2 with 1 d.f..

2.4 Simulation study

Performance of the statistics for detection of over-dispersion in one-way layout of count data with 2 groups are compared by generating data using the IMSL subroutines. We list all the test statistics developed earlier

T_1= C(α) test

T_2= T_1 after restricting the over-dispersion

T_3= Standardised T_1

T_4= T_3 after restricting the over-dispersion

T_5= T_C (Collings and Margolin statistic)

T_6= LR test

T_7= Twice the T_6

We consider

(m_1, m_2) = (5,5), (5,10), (10,10), (10,20), (20,20), (20,40),

c = 0.0, 0.01, 0.02, 0.04, 0.05, 0.0625, 0.1, 0.2, 0.25 and

NR = 10, 20

where

m_i = mean of ith group

and

NR = no. of observations taken from each group.

Table 3.1a presents the empirical power and level of significance for combinations of pair of m_i's and c for 2000 independent samples for NR = 10 .Table 3.1b presents similar results for NR = 20.

This choice of the number of samples implies that a deviation of 0.01 or more between the empirical significance level for a given test and a nominal significance level of 0.05 would be statistically significant. For a better presentation, the empirical levels of significance and powers are multiplied by 1000. An empirical level of significance in 3.1a or 3.1b will be considered close to α =0.05 if it falls in the interval

[0.04,0.06]. Since the tabulated values are multiplied by 1000 the interval is [40,60]. Corresponding to the block for c= 0.00 in table 3.1a, we observe that for all the pairs of means the statistics T_2 and T_5 maintain level of significance close to 0.05. Similarly from table 3.1b, it is observed that statistics T_2, T_5, T_7 maintain a level of significance close to 0.05. Fixing one pair of group means and moving vertically downward along the increasing values of c, we compare the powers of the tests. For all statistics except T_5 empirical level and power were based on critical values of the asymptotic distribution. For T_5 , we first calculated critical values empirically based on 20,000 replications. Using these critical values we then calculated empirical level and power. It is interesting to note that the power of all the tests increases as group means or sample sizes increase. Further T_2 and T_5 hold levels well and they do not differ significantly in power. However, T_5 involves a complicated procedure for calculations of critical values whereas these are readily available for T_2. Thus based on ease of calculation we recommend T_2.

Table 3.1a: 10^3x Empirical power corresponding to α = 0.05 based on 2000 replications. For k = 2 groups and $n_1=n_2$= 10; In each block, rows 1,...,7 correspond to $T_1, T_2, \ldots, T_7$.

(m_1, m_2)

c	(5,5)	(5,10)	(10,10)	(10,20)	(20,20)	(20,40)
0.00	34	41	37	39	37	44
	50	51	49	50	48	52
	67	66	69	72	64	69
	81	80	81	86	82	84
	55	54	54	53	52	53
	11	13	11	15	12	14
	37	39	38	40	40	42
0.01	51	65	71	102	124	201
	68	80	91	119	145	222
	86	110	125	151	191	283
	105	121	149	166	215	307
	75	88	100	127	164	241
	15	22	21	38	48	92
	56	64	76	108	135	209
0.02	73	103	121	189	256	432
	92	115	146	211	286	454
	116	150	200	272	343	524
	137	165	222	298	372	545
	100	129	164	234	308	476
	22	34	47	86	132	258
	82	103	130	200	265	439
0.04	119	189	255	420	527	752
	145	209	283	448	567	767
	198	275	345	520	637	815
	219	301	365	536	664	823
	171	232	306	474	539	786
	50	92	129	256	372	616
	132	194	266	430	534	760

(contd.)

Table 3.1a (contd.)

0.05	153	249	316	524	655	836
	178	268	346	546	685	845
	228	332	424	611	735	882
	254	364	444	625	758	887
	202	291	378	568	699	860
	68	124	183	350	483	738
	162	255	333	535	666	846
0.06	198	316	404	630	764	904
	228	345	435	651	783	910
	275	416	509	711	822	930
	297	436	504	729	841	934
	248	376	465	671	790	968
	89	176	258	479	628	889
	203	331	417	642	770	910
0.10	318	514	658	831	924	974
	343	536	681	841	932	976
	407	613	747	879	951	936
	439	628	764	885	954	987
	374	566	708	859	937	979
	188	351	478	730	864	954
	330	519	668	834	928	977
0.20	642	820	919	973	994	999
	660	833	928	974	994	999
	739	873	944	982	997	1000
	754	880	949	982	997	1000
	704	855	973	979	996	1000
	467	717	853	955	984	999
	668	832	925	976	995	999
0.25	734	897	957	989	999	1000
	753	903	961	989	999	1000
	804	920	971	993	1000	1000
	821	926	973	993	1000	1000
	785	912	965	990	999	1000
	598	817	919	991	996	1000
	752	904	960	993	1000	1000

Table 3.1b : 10^3 x empirical power corresponding to $\alpha = 0.05$ based on 2000 replications. for k = 2 groups and $n_1 = n_2 = 20$; in each block rows 1,...,7 correspond to $T_1, T_2, \ldots, T_7$.

(m_1, m_2)

c	(5,5)	(5,10)	(10,10)	(10,20)	(20,20)	(20,40)
0.00	35	42	40	42	41	44
	47	53	51	52	56	54
	53	60	57	63	62	65
	67	76	72	80	77	79
	48	50	51	51	53	53
	4	9	12	10	11	11
	46	48	45	52	51	51
0.01	64	77	96	139	177	313
	82	88	117	154	201	344
	96	116	131	192	239	385
	112	130	151	210	260	410
	82	95	118	159	205	354
	19	24	36	63	90	168
	74	91	107	158	200	339
0.02	96	140	169	309	409	661
	116	155	195	337	437	679
	135	185	236	382	483	719
	155	199	255	405	506	732
	118	163	210	345	447	694
	38	58	86	169	266	503
	109	152	192	335	443	685
0.05	228	399	525	778	893	982
	252	423	551	788	901	983
	287	471	588	820	918	988
	313	494	613	878	925	989
	260	434	567	798	904	986
	122	242	362	647	796	961
	253	422	555	796	903	986

(contd.)

Table 3.1b (contd.)

0.06	298	514	645	870	950	996
	323	538	669	873	953	996
	365	581	709	902	967	998
	388	599	728	905	968	998
	334	550	688	881	961	997
	167	349	493	776	906	989
	379	537	676	883	960	997
0.10	506	762	880	980	999	1000
	531	776	887	982	999	1000
	586	808	917	986	1000	1000
	609	819	922	987	1000	1000
	550	786	905	984	999	1000
	353	635	793	956	992	1000
	544	783	901	985	999	1000
0.20	867	976	996	1000	1000	1000
	877	978	996	1000	1000	1000
	904	983	999	1000	1000	1000
	912	984	999	1000	1000	1000
	889	978	998	1000	1000	1000
	772	951	989	1000	1000	1000
	886	980	999	1000	1000	1000
0.25	939	995	1000	1000	1000	1000
	944	995	1000	1000	1000	1000
	954	997	1000	1000	1000	1000
	959	997	1000	1000	1000	1000
	950	996	1000	1000	1000	1000
	883	983	1000	1000	1000	1000
	949	997	1000	1000	1000	1000

Example : Collings & Margolin(1985) presented the summary statistics for the data resulting from an experiment where 4-nitro-ortho-phenylenediamine (4 NoP) was tested in a standard Ames test with Salmonella Strain TA98 and without metabolic activation. Dosage levels of 4 NoP were set at 0.0, 0.3, 1.0, 3.0, 10.0 μg/plate and 20 replicate observations for each dose was collected. Summary statistics are presented in table 3.1.

Table 3.2: Summary Statistics of an Ames test for 4 NoP

Dose of 4 NoP(μg/plate)

statistic	0.0	0.3	1.0	3.0	10.0
sample mean	17.8	51.7	110.9	283.5	692.3
s_i^2 sample variance	17.5	81.0	175.4	1131.5	4584.4
S^2 biased samp.var	16.6	77.0	166.6	1074.9	4355.2

First, we calculate T_c given by Collings and Margolin(1985) which has the value 492.155.This value of T_s is to be compared with the percentage point of the r.v.

$T = \Sigma w_i \chi^2(19)$,

where $w_i = \bar{y}_i / \bar{y}$, $i = 1,..,5$

and $\chi^2(19)$ = r.v. distributed as χ^2 with 19 d.f..

As the percentage-point of T is not readily available , we approximate the distribution of T by $g\chi^2(h)$. The unknown values of g and h are obtained by equating the first two moments to those of $g\chi^2(h)$.

Thus

$$gh = 95$$

$$2g^2h = 38.\Sigma w_i^2 = 408.583$$

Appealing to the normal approximation for χ^2 distribution with large d.f.

$$z = (492.155 - 95)/\sqrt{408.583} = 19.65,$$

which is highly significant.

The hypothesis of absence of negative binomial variation is rejected strongly.

The ordinary C(α) test T_1 in (3.3.18) and range-justified C(α) test T_2 in (3.3.21) yield 18.909 and 18.914 respectively, again showing strong evidence of presence of negative binomial variation. Similarly T_3 in (3.3.20) and T_4 in (3.3.22) both yield 19.65. This value is equal to the value of z obtained earlier. Thus the simpler methods presented in this chapter lead to the same conclusions as the statistic T_c.

CHAPTER 4

ANALYSIS OF ONE-WAY LAY-OUT OF COUNT DATA WITH NEGATIVE BINOMIAL VARIATION WHEN THE DISPERSION PARAMETERS ARE EQUAL

4.1 Introduction

In the last chapter, we developed procedures for the analysis of one-way layout of count data assuming the underlying distribution to be Poisson. Methods for detecting extra-poisson variation (described by negative binomial distribution) were also developed and discussed . This chapter presents different methods of comparing means of several negative binomial distributions having a common dispersion parameter. As described earlier, this analysis parallels the usual ANOVA of one-way layout of data with the underlying distribution being normal distribution with equal variances.

Assume $y_{ij} \sim NB(m_i, c)$, $\quad i = 1,..,k$, $\quad j = 1,..,n_i$, (4.1.1)

as data coming from an experiment designed to test equality of k group means. The probability mass function of y_{ij} is given by

$$P(y_{ij} = y) = \frac{\Gamma(y + c^{-1})}{y!\Gamma(c^{-1})}\left(\frac{cm_i}{1 + cm_i}\right)^{y}\left(\frac{1}{1 + cm_i}\right)^{c^{-1}} . \qquad (4.1.2)$$

The competing hypotheses are

$$H_0: m_1 = m_2 = \ldots = m_k \qquad (4.1.3)$$

and

H_A: at least one inequality among m_i's

while c is the unknown dispersion parameter

both under H_0 & H_A.

The methods presented are

i) The likelihood ratio test(discussed in section 4.2) ,

ii) C(α) tests (discussed in section 4.3) ,

and

iii) Tests based on variance stabilising transformation of negative binomial variable (discussed in section 4.4) .

All these methods have been developed assuming a common c. However, in a practical context this assumption of homogeneity of dispersion parameters might not be tenable, so we also develop methods for testing the assumption of common c. The methods used for this are

i) LR test for the assumption of common c (discussed in section 4.6) and

ii) C(α) tests for the assumption of common c (discussed in section 4.6).

In sections 4.2-4.5 procedures for testing equality of means in the presence of a common c are developed and studied. In section 4.6 C(α) statistics for testing of homogeneity of dispersion parameters are developed.

4.2 The Likelihood Ratio Test

Denote the log-likelihood of the data excepting a constant(depending on the data only) under the alternative hypothesis by ℓ. The likelihood ratio test is defined as

$$\chi^2(a) = 2(\hat{\ell}_1 - \hat{\ell}_0), \qquad (4.2.1)$$

where $\hat{\ell}_1$ and $\hat{\ell}_0$ are the estimated log-likelihood under the null and the alternative hypotheses; a is the difference of the number of parameters under the null and the alternative hypotheses.

Under H_A the log-likelihood ℓ is given by,

$$\ell = \sum_i \sum_j \left(\sum_{t=0}^{y_{ij}-1} \ln(1+ct) \right) + \sum_i y_{i+} (\ln m_i - \ln(1+cm_i)$$

$$- \sum_i (n_i/c)\ln(1+cm_i) , \qquad (4.2.2)$$

for $i = 1,\ldots,k$
and $j = 1,\ldots,n_i$.

Under H_0, the common mean m and common dispersion parameter c are unknown.The maximum likelihood estimate (mle) of m is obtained by solving

$$\frac{\partial \ell}{\partial m} = \sum_{i=1}^{k} y_{i+} \left(\frac{1}{m} - \frac{c}{1+cm} \right) - \sum_i \frac{n_i}{c} \cdot \frac{c}{1+cm} = 0.$$

After simplification , we find

$$\hat{m} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \Big/ \sum_{i} n_i = \bar{y} \ . \qquad (4.2.3)$$

The mle of c denoted by c_o is obtained by solving the following equation iteratively :

$$\frac{\partial \ell}{\partial c} = - \sum_{i} \sum_{j} \sum_{t} \frac{1}{c(1 + ct)} + \frac{n.\ln(1+c\bar{y})}{c^2} = 0 \ , \qquad (4.2.4)$$

where $\sum n_i = n$.

Under the alternative hypothesis, k different group means and the common dispersion parameter are unknown.The mle of m_i is obtained from

$$\frac{\partial \ell}{\partial m_i} = y_{i+} \left(\frac{1}{m_i} - \frac{c}{1 + cm_i} \right) - \frac{n_i}{c} \cdot \frac{c}{1 + cm_i} = 0 \ . \qquad (4.2.5)$$

After simplification ,

$$\hat{m}_i = \bar{y}_i \ . \qquad (4.2.6)$$

When the group means are assumed different, the mle of c denoted by c_1 is obtained by solving the following equation iteratively

$$\frac{\partial \ell}{\partial c} = - \sum_{i} \sum_{j} \sum_{t} \frac{1}{c(1+ct)} + \sum_{i} \frac{n_i \ln(1+c\bar{y}_i)}{c^2} = 0. \qquad (4.2.7)$$

Using these mles, the likelihood ratio statistic χ_1^2 is derived as

$$\chi_l^2 = 2(\hat{\ell}_1 - \hat{\ell}_0)$$

$$= 2\sum_{i=1}^{k}\left[n_i.\left[\bar{y}_i \ln\left[\frac{\bar{y}_i(1 + c_0\bar{y})}{\bar{y}(1 + c_1\bar{y}_i)}\right]\right.\right.$$

$$\left.\left. + \frac{\ln(1+c_0\bar{y})}{c_0} - \frac{\ln(1+c_1\bar{y}_i)}{c_1}\right]\right]$$

$$+2\sum_{i}\sum_{j=1}^{n_i}\sum_{t=0}^{y_{ij}-1} \ln\left(\frac{1 + c_1 t}{1 + c_0 t}\right) . \qquad (4.2.8)$$

The statistic χ_l^2 has a complicated expression involving two estimates of c. Both require solving complicated non-linear equations using iterative procedures. Furthermore, this does not provide much information regarding the contribution of individual group means to χ_l^2.

4.3 C(α) tests

In this section , C(α) tests based on the method developed in chapter 2 are derived. Reparametrise the m_i's as

$m_i = m + \phi_i$, $i = 1,\ldots,k$,

s.t. $\phi_k = 0$. (4.3.1)

Thus the usual competing hypotheses in (4.1.3) reduces to

H_0: $\phi_1 = \phi_2 = \ldots = \phi_k$, m,c being unknown,

and

H_A: at least one inequality among ϕ_i's,

ϕ_i's and m, c being unknown. (4.3.2)

Define $\Phi=(\phi_1,\ldots,\phi_{k-1})'$ and $\theta=(\theta_1,\theta_2)'=(m,c)'$

$$\psi_i(\theta) = \left.\frac{\partial \ell}{\partial \phi_i}\right]_{\Phi = 0}, \quad i = 1,\ldots,k-1,$$

$$\gamma_i(\theta) = \left.\frac{\partial \ell}{\partial \theta_i}\right]_{\Phi = 0}, \quad i = 1,2, \qquad (4.3.3)$$

$$D_{ij}(\theta) = E\left[\left.\frac{\partial^2 \ell}{\partial \phi_i \partial \phi_j}\right]_{\Phi = 0}\right], \quad 1 \le i,j \le k-1,$$

$$A_{ij}(\theta) = E\left[\left.\frac{\partial^2 \ell}{\partial \phi_i \partial \theta_j}\right]_{\Phi = 0}\right], \quad \begin{array}{l} 1 \le i \le k-1; \\ j = 1,2, \end{array}$$

$$D_{ij}(\theta) = E\left[\left.\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}\right]_{\Phi = 0}\right], \quad 1 \le i,j \le 2,$$

(4.3.4)

where ℓ denotes the log-likelihood in terms of Φ, m, c.

Now,

$$\ell = \sum_i \sum_j \left[\sum_{t=0}^{y_{ij}-1} \ln(1 + ct)\right] - \sum_i \frac{n_i}{c} \ln(1 + c(m + \phi_i))$$

$$+ \sum_i y_{i+}(\ln(m + \phi_i) - \ln(1 + c(m + \phi_i))), \qquad (4.3.5)$$

and we have

$$\psi_i = \frac{n_i(\bar{y}_i - m)}{m(1 + cm)}, \qquad i = 1,\ldots,k-1,$$

$$\gamma_1 = \sum_{i=1}^{k} \frac{n_i(\bar{y}_i - m)}{m(1 + cm)}, \qquad (4.3.6)$$

$$\gamma_2 = \frac{n.\ln(1 + cm)}{c^2} - \sum_i \sum_j \sum_t \frac{1}{c(1 + ct)} \quad ,$$

$$D_{ij} = \frac{n_i}{m(1 + cm)} \quad , \qquad 1 \le i = j \le k-1,$$

$$= \quad 0 \quad , \qquad \text{otherwise,}$$

$$A_{ij} = \frac{n_i}{m(1 + cm)} \quad , \qquad 1 \le i \le k-1;\ j = 1,$$

$$= \quad 0 \quad , \qquad 1 \le i \le k-1;\ j = 2,$$

$$B_{11} = \frac{n}{m(1 + cm)} \quad ,$$

$$B_{22} = \quad b \ \text{(say)},$$

$$B_{12} = B_{21} = 0 \ . \qquad (4.3.7)$$

The $C(\alpha)$ test is constructed using the following adjusted scores S_i, as discussed in chapter 2,

$$S_i = \psi_i - \beta_{i1}\gamma_1 - \beta_{i2}\gamma_2 \quad , \qquad (4.3.8)$$

$$i = 1,\ldots,k-1 \ .$$

The dispersion matrix of $(\psi_1,\psi_2,\ldots,\psi_{k-1},\gamma_1,\gamma_2)$ is given by V with the following structure,

$$V = \begin{bmatrix} D & A \\ A' & B \end{bmatrix} \quad , \qquad (4.3.9)$$

where

D is $(k-1)\times(k-1)$, A is $(k-1)\times 2$ and B is 2×2 matrices.

For $k = 3$, the matrix V has the form

$$V = \frac{1}{m(1 + cm)} \begin{bmatrix} n_1 & 0 & n_1 & 0 \\ 0 & n_2 & n_2 & 0 \\ n_1 & n_2 & n & 0 \\ 0 & 0 & 0 & d \end{bmatrix} .$$

The regression coefficient calculated from V are

$$\beta_{i1} = \frac{n_i}{n} \quad \text{and}$$

$$\beta_{i2} = 0 \qquad (4.3.10)$$

$$i=1,\ldots,k-1.$$

Substituting the values of β_{ij}'s from (4.3.10) in S_i's in (4.3.8), we have

$$S_i = \psi_i - \beta_i \gamma_1 \ , \qquad (4.3.11)$$

where $\beta_i = \frac{n_i}{n}$.

The var-covar matrix of $(S_1, S_2, .., S_{k-1})$ is given by

$$V_{11.2} = D - AB^{-1}A' .$$

Define

$$d_i = \frac{n_i}{m(1 + mc)} \ , \qquad \text{for } i = 1,\ldots,k$$

and a vector $d' = (d_1,\ldots,d_{k-1})$. (4.3.12)

Using the values of d_i's, we have

$$V_{11.2} = \text{Diag}(d) - \frac{dd'}{1'd + d_k}$$

and

$$V_{11.2}^{-1} = \text{Diag}(1/d_1, 1/d_2, \ldots, 1/d_{k-1})$$

$$- \frac{11'}{d_k} . \qquad (4.3.13)$$

Thus the C(α) statistic ,

$$\chi_C^2 = (S_1, \ldots, S_{k-1}) V_{11.2}^{-1} (S_1, \ldots, S_{k-1})' \qquad (4.3.14)$$

follows $\chi^2(k-1)$, appealing to asymptotic normality of scores. After some simplification, we obtain

$$\chi_C^2 = \sum_{i=1}^{k-1} \frac{S_{i1}^2}{d_i} + \frac{(\sum_{i=1}^{k-1} S_i)^2}{d_k} . \qquad (4.3.15)$$

From (4.3.6), define

$$\psi_k = \frac{n_k (\bar{y}_k - m)}{m(1 + cm)} .$$

From (4.3.11), define

$$\beta_k = \frac{n_k}{n} .$$

Again, from (4.3.6) and definitions of ψ_k and β_k ,

$$\sum_{i=1}^{k-1} S_i = \sum_{i=1}^{k-1} \psi_i - \gamma_1 \sum_{i=1}^{k-1} \beta_i ,$$

$$= -(\psi_k - \beta_k \gamma_1)$$

$$= - S_k . \qquad (4.3.16)$$

Using, (4.3.16), in (4.3.15) , we get,

$$\chi_C^2 = \sum_{i=1}^{k} \frac{S_i^2}{d_i} \quad . \qquad (4.3.17)$$

χ_C^2 in (4.3.17) involves two nuisance parameters m and c. These parameters will be replaced by their $\sqrt{n}$-consistent estimates to obtain a numerical value which will in turn be compared with percentage points of χ^2 with (k-1) d.f. The two sets of $\sqrt{n}$-consistent estmates are

i) maximum likelihood estimates of m & c which are $\bar{y}$ & c_o respectively,

ii) estimates of m & c using method of moments are $\bar{y}$ & c' respectively, where

$$c' = \frac{s^2 - \bar{y}}{\bar{y}^2}$$

and

$$s^2 = \frac{1}{n-1} \sum_i \sum_j (y_{ij} - \bar{y})^2 .$$

When m is replaced by $\bar{y}$ in (4.3.6), we get

$$\gamma_1 = 0.$$

This simplifies S_i as

$$S_i(\hat{\theta}) = \psi_i(\hat{\theta}) \quad . \qquad (4.3.18)$$

With this simplification, after some algebra χ_C^2 reduces to the elegant forms ,

a) $$\chi^2_{C(m)} = \sum_{i=1}^{k} \frac{n_i (\bar{y}_i - \bar{y})^2}{\bar{y}(1 + c_o \bar{y})} \qquad (4.3.19)$$

and

b) $$\chi^2_{C(mm)} = \sum_{i=1}^{k} \frac{n_i (\bar{y}_i - \bar{y})^2}{\bar{y}(1 + c' \bar{y})} \qquad (4.3.20)$$

When k = 2, the square root of $\chi^2_{C(m)}$ is identical to the formula

$$z = \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \cdot \frac{(\bar{y}_1 - \bar{y}_2)}{\sqrt{\bar{y} (1 + c_o \bar{y})}}$$

derived by the method given in Moran(1970, p 51). The statisitic $\chi^2_{C(m)}$ has a form similar to the test statistic derived by Margolin(1985) for testing the trend in negative binomial means in presence of a common dispersion parameter. Substituting the value of c' in (4.3.20), $\chi^2_{C(mm)}$ reduces to the CATANOVA (categorical analysis of variance) statistic in Light and Margolin(1971).

The form of $\chi^2_{C(m)}$ and $\chi^2_{C(mm)}$ provides the similar interpretation as the normal theory ANOVA in data-analysis. The numerator may be considered as the SS due to difference among groups. Using the above fact, it is possible to measure the contribution of each group towards the variability among the group means.

4.4 Test statistics based on variance stabilising transformations

Anscombe(1948) suggests two asymptotically variance stabilising transformations for negative binomial distribution:

When $Y \sim NB(m,c)$

$$\text{i. } X_1 = \sqrt{1/c - 0.5}\,\sinh^{-1}\left[\frac{Y + 0.375}{1/c - 0.75}\right]^{\frac{1}{2}} \qquad (4.4.1)$$

$$\text{ii. } X_2 = \sqrt{1/c - 0.5}\,\ln\left(Y + \frac{1}{2c}\right) \qquad (4.4.2)$$

In the case when c is unknown , c is replaced by its mle in (4.4.1) and (4.4.2) and then we proceed to analyse the transformed data using the normal theory ANOVA.

C(α) test for one-way layout of data having normal variation:

Let $x_{ij} \sim N(\mu_i,\sigma^2)$ for $i = 1,\ldots,k; j = 1,\ldots,n_i$

The competing hypotheses are

H_o: $\mu_1 = \mu_2 = \ldots = \mu_k$

and

H_A: at least one strict inequality among μ_i's (4.4.3)

σ^2 being a nuisance parameter both under H_o and H_A.

Reparametrise $\mu_i = \mu + \phi_i$, $i = 1,\ldots,k$

with $\phi_k = 0$. (4.4.4)

For this data set the log-likelihood ℓ excepting a constant (solely depending on the data) is given by

$$-2\ell = n.\ln\sigma^2 + \sum_i \sum_j \frac{(x_{ij} - \mu_i)^2}{\sigma^2} \quad . \qquad (4.4.5)$$

Using the same notation as in the previous section and replacing $\theta' = (\mu, \sigma^2)$;we obtain

$$\psi_i(\theta) = \frac{n_i(\bar{x}_i - \mu)}{\sigma^2}, \quad i = 1,..,k-1$$

$$\gamma_1(\theta) = \sum_i \frac{n_i(\bar{x}_i - \mu)}{\sigma^2},$$

$$\gamma_2(\theta) = \frac{n}{2\sigma^4}\left[\frac{\sum_i \sum_j (x_{ij} - \mu)^2}{n} - \sigma^2\right], \qquad (4.4.6)$$

$$D_{ij}(\theta) = \frac{n_i}{\sigma^2}, \qquad 1 \le i = j \le k-1,$$

$$= 0, \qquad \text{otherwise},$$

$$A_{ij}(\theta) = \frac{n_i}{\sigma^2}, \qquad 1 \le i \le k-1;\ j = 1,$$

$$= 0, \qquad \text{otherwise},$$

$$B_{11}(\theta) = \frac{n}{\sigma^2},$$

$$B_{22}(\theta) = \frac{n}{2\sigma^4},$$

$$B_{12}(\theta) = B_{21}(\theta) = 0, \qquad (4.4.7)$$

and $n = \sum_i n_i$.

In this case also, the matrix V and the efficient scores have the same form as before but m(1+cm) is replaced by σ^2. Again, defining

$$S_i = \psi_i(\theta) - \beta_{i1}\gamma_1(\theta) - \beta_{i2}\gamma_2(\theta), \qquad (4.4.8)$$

for $i = 1,\ldots,k-1$,

the test statistic is

$$\chi_T^2 = S'(D - AB^{-1}A')^{-1}S \ , \qquad (4.4.9)$$

where $S' = (S_1,\ldots,S_{k-1})'$

Since χ_T^2 involves nuisance parameters μ , σ^2, we replace them by their mles $\bar{x}$ and S^2. After some algebra, we find

$$\chi_T^2 = \frac{\sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2}{S^2} \ , \qquad (4.4.10)$$

where

$$\bar{x} = \frac{\sum_i n_i \bar{x}_i}{n} \ ,$$

and

$$S^2 = \frac{1}{n} \sum_i \sum_j (x_{ij} - \bar{x})^2 \ .$$

Thus

$$\chi_T^2 = \frac{n \sum_i n_i (\bar{x}_i - \bar{x})^2}{\sum_i \sum_j (x_{ij} - \bar{x})^2 + \sum_i n_i (\bar{x}_i - \bar{x})^2} \ .$$

Let $F = \dfrac{(n-k)\sum_i n_i (\bar{x}_i - \bar{x})}{(k-1)\sum_i \sum_j (x_{ij} - \bar{x}_i)^2}$ then

$$\chi_T^2 = \frac{n.F}{(n-k)/(k-1) + F} \ .$$

It could easily be seen that

i) χ_T^2 is a multiple of a random variable having a beta distribution, hence the distribution of χ_T^2 can be used to

compute the required percentage points exactly ,

ii) χ_T^2 is a monotonically increasing function of F, where F is the usual F-statistic used in the analysis of one-way data with normal variation ,

iii)For large n , χ_T^2 is approximately distributed as $\chi^2(k-1)$, which agrees with the asymptotic distribution of F.This illustrates another method of deriving the well-known ANOVA F-statistic. For the purpose of comparing the means of negative binomial distributions ,we next show that, for large m and small c, X_2 is a linear function of X_1.

Lemma:The transformations $X_1$and $X_2$are linearly related for large values of m and/or small values of c.

Proof :It is sufficient to show that

$$X_1 = aX_2 + b + \text{negligible terms}$$

where a, b do not depend upon $Y \sim NB(m,c)$.

$$X_1 = 2\sqrt{\frac{1}{c} - \frac{1}{2}}\ \sinh^{-1}\sqrt{\frac{Y + 0.375}{1/c - 0.75}}$$

$$= 2k_1 \ln\left(\sqrt{Y + 0.375} + \sqrt{Y + \frac{1}{c} - 0.375}\right) - k_2$$

$$= k_1 \ln\left(Y + \frac{1}{2c}\right) - k_2 + 2k_1 \ln\left(\sqrt{1 + Ay} - \sqrt{1 - Ay}\right)$$

$$= X_2 - k_2 + k_1 \ln 2 + k_1 O\left(\frac{1}{Y^2}\right),$$

where

$$Ay = \frac{0.375 - \frac{1}{2c}}{Y + \frac{1}{2c}}$$ and k_1 and k_2 are constants.

Thus we can conclude that the F-statistics using X_1 and X_2 will produce approximately equal values as F-statistic is invariant under scale and location transformation.

4.5 Simulation Studies

Performances of the statistics developed earlier are compared by conducting a simulation study. We consider the problem of testing of equality of means of two and three negative binomial distributions with common dispersion parameter. First we study the behaviour of the test statistics in maintaining the pre-assigned level of significance $\alpha(= 0.05)$. The test statistics under study are

χ_1^2 = Likelihood ratio test,

$\chi_{C(m)}^2$ = C(α) test using maximum likelihood estimates,

$\chi_{C(mm)}^2$ = C(α) test using method of moments estimates

and

T_2 = Test statistic obtained using the variance-stabilising transformation .

Empirical level of significance was based on 2000 replications. In each replication, two samples of size NR are generated from NB(m,c) and the above statistics are calculated.

Table 4.1a presents the proportions of rejections

multiplied by 1000 for the following set of parameters

$m = 5, 10, 20, 50,$

$NR = 5, 10, 20,$

$c = 0.05, 0.25, 0.50.$

Table 4.1b presents empirical levels of significance multiplied by 1000 based on 2000 replications . In each replication three independent samples of size NR are drawn from NB(m,c) and the above statistics are calculated. The values of the parameters are the same as the ones used for table 4.1a. The percentage points at 5 % level are used in both the cases. For $\alpha = 0.05$, the empirical level of significance will be considered close to α if it falls within the interval [0.04,0.06]. Since we have multiplied the empirical levels of significance by 1000, the interval changes to [40,60]. From table 4.1a, we find that a lot of times the empirical level of significance of χ_1^2 falls outside the interval(actually exceeding the upper limit 60) suggesting that LR test is a liberal test. The statistic $\chi^2_{C(mm)}$ performs poorly for small sample size(=5). The $\chi^2_{C(m)}$ and T_2 maintain level of significance quite well. From table 4.1b, we observe that LR test once again performs poorly excepting when m and NR are large. $\chi^2_{C(mm)}$ performs poorly for small and moderate sample sizes and small values of m(=5). The statistics $\chi^2_{C(m)}$ and T_2 maintain levels of significance well.

Table 4.1a

10^3x empirical levels; α = 0.05; based on 2000 replications In each block rows 1, 2, 3, 4 correspond to statistic $\chi^2_{1\&}$ $\chi^2_{C(m)}$, $\chi^2_{C(mm)}$ and T_2 respectively for 2 groups.

m	$n_1 = n_2 = 5$			$n_1 = n_2 = 10$			$n_1 = n_2 = 20$		
	c			c			c		
	0.05	0.25	0.50	0.05	0.25	0.50	0.05	0.25	0.50
5	92	98	97	64	64	63	84	61	59
	59	52	43	44	49	45	57	58	50
	39	39	32	40	44	41	55	55	51
	28	49	47	30	45	44	40	59	56
10	97	95	95	58	60	63	68	59	60
	60	52	46	47	47	45	57	55	53
	38	38	35	41	41	41	53	52	53
	34	51	49	34	46	44	52	59	58
20	95	94	95	60	63	61	63	60	57
	59	52	43	47	46	47	58	55	54
	40	38	33	42	41	39	56	51	54
	46	52	50	42	42	40	57	58	58
50	90	94	89	65	65	64	62	63	63
	61	52	43	51	49	50	59	55	55
	41	39	34	42	42	42	56	53	53
	51	51	48	49	48	47	58	55	60

Table 4.1b

10^3x empirical levels: α = 0.05; based on 2000 replications. In each block, rows 1, 2, 3, 4 correspond to statistics χ^2_1, $\chi^2_{C(m)}$, $\chi^2_{C(mm)}$, T_2 respectively , for three groups

m	$n_1 = n_2 = n_3 = 5$			$n_1 = n_2 = n_3 = 10$			$n_1 = n_2 = n_3 = 20$		
	c			c			c		
	0.05	0.25	0.50	0.05	0.25	0.50	0.05	0.25	0.50
5	91	93	94	75	62	65	82	68	63
	41	41	34	48	40	39	53	54	49
	29	30	27	41	36	37	49	50	49
	29	44	47	38	50	53	39	54	50
10	91	93	97	66	59	64	62	62	61
	46	40	34	52	44	39	54	50	49
	33	31	28	45	40	35	51	47	48
	37	49	50	46	49	50	47	50	45
20	93	95	100	68	61	65	62	61	62
	45	41	38	45	43	40	52	52	48
	32	30	31	40	42	33	51	50	51
	42	47	47	50	50	49	49	50	51
50	91	96	98	67	61	62	56	53	56
	43	40	40	48	45	42	48	45	45
	31	31	32	43	45	48	47	46	44
	48	47	45	52	52	52	46	46	44

After studying the behaviour of empirical levels of significance, we study the performance of these statistics with respect to power. We consider the following values of the parameters. For two groups,

$m_1 = m$,

$m_2 = m + \phi$ and $\delta = \phi / m$,

where

$m = 5, 10, 20, 50$,

$\delta = 0.0, 0.2, 0.4, 0.6, 0.8$,

$c = 0.05, 0.25$

and

$NR = 10, 20$.

For three groups ,

$m_1 = m$,

$m_2 = m+\phi_2$ and $\delta_2 = 10\ \phi_2 / m$,

$m_3 = m+\phi_3$ and $\delta_3 = 10\ \phi_3 / m$,

$\delta = (\delta_2, \delta_3)$,

where

$m = 5, 10, 20, 50$,

$\delta = (0,0), (0,2), (2,4), (4,6), (6,8)$,

$NR = 10, 20$

and

$c = 0.05, 0.25$.

Table 4.2a presents the values of power for two groups

when the common dispersion parameter is 0.05. Empirical levels of significance for $\chi^2_{C(m)}$ is always higher than those of T_2 but always less than 50. We alsc see that the power of $\chi^2_{C(m)}$ is higher than those of T_2. Table 4.2b also indicates that $\chi^2_{C(m)}$ is better than its competitors in terms of power. In both the tables, we observe that a significant increase in power is acheived by increasing the common sample size from 10 to 20.

Table 4.3a and 4.3b presents result of simulation study for three groups. In table 4.3a, for small values of m, T_2 χ^2_L perform poorly. For moderately large and large values of m and small values of c , $\chi^2_{C(m)}$ performs better than T_2 and $\chi^2_{C(mm)}$.For c = 0.25 and NR = 10 even though T_2, $\chi^2_{C(m)}$, $\chi^2_{C(mm}$ have their empirical levels of significance close to 0.05, power of $\chi^2_{C(m)}$ seems to be better than the other two. Based on the above simulation study $\chi^2_{C(m)}$ is recommended for the analysis of one-way layout of count data showing extra-poisson variation represented by negative binomial distribution.

Table 4.2a

10^3x Empirical power corresponding to nominal significance level $\alpha = 0.05$, based on 2000 replications for 2 groups.

$m_1 = m$, $m_2 = m + \phi$, $\delta = \phi / m$

$c = 0.05$

	$n_1 = n_2 = 10$					$n_1 = n_2 = 20$				
m	δ					δ				
	0.0	0.2	0.4	0.6	0.8	0.0	0.2	0.4	0.6	0.8
5	64	148	370	640	837	84	236	621	895	985
	44	117	327	589	805	57	212	600	882	982
	40	106	299	564	781	55	206	588	877	980
	30	92	272	532	759	40	179	547	857	972
10	58	200	544	835	962	68	344	824	983	1000
	47	171	489	801	948	57	327	812	982	1000
	41	157	468	781	936	53	312	806	981	1000
	34	150	453	781	933	52	304	788	979	999
20	60	274	705	941	993	63	465	935	998	1000
	47	238	660	920	987	58	443	927	997	1000
	42	214	635	903	985	56	432	921	997	1000
	42	220	647	908	985	57	435	916	996	1000
50	65	350	831	985	1000	62	600	985	1000	1000
	51	313	804	979	1000	59	579	984	1000	1000
	42	290	776	974	998	56	565	980	1000	1000
	49	302	787	971	999	58	556	982	1000	1000

Table 4.2b

10^3x Empirical power corresponding to nominal significance level $\alpha = 0.05$, based on 2000 replications for 2 groups.

$m_1 = m$, $m_2 = m + \phi$, $\delta = \phi / m$

$c = 0.25$

m	$n_1 = n_2 = 10$					$n_1 = n_2 = 20$				
	δ					δ				
	0.0	0.2	0.4	0.6	0.8	0.0	0.2	0.4	0.6	0.8
5	64	101	216	377	546	61	147	381	655	827
	49	82	180	320	481	58	132	358	626	808
	44	72	165	300	452	55	132	344	610	798
	45	78	175	301	455	59	126	323	561	757
10	60	112	269	446	633	59	176	454	737	899
	47	92	215	388	572	55	159	430	711	881
	41	81	197	359	540	52	154	420	698	874
	46	88	204	367	534	59	146	389	656	842
20	63	120	296	507	694	60	191	497	783	928
	46	98	253	441	631	55	178	473	758	921
	41	91	220	401	595	51	174	463	747	909
	42	95	237	411	593	58	160	438	708	882
50	65	140	315	532	732	63	200	537	822	951
	49	110	274	467	681	55	184	502	797	940
	42	95	241	431	622	53	184	491	786	929
	48	104	249	439	626	55	163	465	744	912

Table 4.3a

10^3x Empirical power corresponding to nominal significance level $\alpha = 0.05$, based on 2000 replications for 3 groups.

$m_1 = m$, $m_2 = m + \phi_2$, $m_3 = m + \phi_3$;

$\delta_2 = 10x\phi_2/m$, $\delta_3 = 10x\phi_3/m$; $\delta = (\delta_2, \delta_3)$.

$c = 0.05$

m	$n_1 = n_2 = n_3 = 10$					$n_1 = n_2 = n_3 = 20$				
	δ (0,0)	(0,2)	(2,4)	(4,6)	(4,8)	δ (0,0)	(0,2)	(2,4)	(4,6)	(6,8)
5	75	147	303	542	754	82	247	524	848	968
	48	114	241	465	689	53	221	488	825	960
	41	104	222	440	664	49	218	481	821	956
	38	94	222	453	672	39	184	433	793	946
10	66	201	437	773	928	62	361	741	972	999
	52	160	388	700	990	54	337	716	965	999
	45	143	364	677	871	51	335	709	963	998
	46	149	372	708	885	47	314	698	959	997
20	68	278	590	901	984	62	488	884	999	1000
	45	226	524	851	974	52	470	866	998	1000
	40	212	501	843	971	51	464	861	998	1000
	50	218	529	865	970	49	442	855	998	1000
50	67	356	741	971	999	56	636	969	1000	1000
	48	309	684	952	997	48	617	961	1000	1000
	43	289	651	942	996	47	611	958	1000	1000
	52	296	689	954	996	46	585	956	1000	1000

Table 4.3b

10^3x Empirical power corresponding to nominal significance level $\alpha = 0.05$, based on 2000 replications for 2 groups.

$m_1 = m$, $m_2 = m + \phi_2$, $m_3 = m + \phi_3$;

$\delta_2 = 10x\phi_2/m$, $\delta_3 = 10x\phi_3/m$; $\delta = (\delta_2, \delta_3)$.

$c = 0.25$

$n_1 = n_2 = n_3 = 10$ $\qquad$ $n_1 = n_2 = n_3 = 20$

m	δ (0,0)	(0,2)	(2,4)	(4,6)	(6,8)	δ (0,0)	(0,2)	(2,4)	(4,6)	(6,8)
5	62	108	174	310	440	68	160	304	551	746
	40	79	129	229	363	54	137	270	497	709
	36	72	123	212	336	50	138	268	491	698
	50	83	141	250	359	54	128	251	469	655
10	59	180	290	369	521	63	183	400	653	834
	44	91	160	283	439	50	164	374	604	806
	40	82	143	264	408	47	160	324	593	794
	49	85	166	300	435	50	138	301	567	751
20	61	126	230	413	570	61	199	428	711	881
	43	97	179	327	499	52	177	394	667	859
	42	89	159	296	459	50	176	366	653	846
	50	95	189	336	489	50	154	336	631	807
50	61	127	238	441	620	53	200	428	743	907
	45	104	191	351	538	45	184	394	604	890
	45	88	165	319	494	46	176	381	683	880
	52	92	197	362	520	46	159	355	674	841

4.6 Testing The Homogeneity Of Dispersion Parameters Of Several Negative Binomial Distributions

Testing of the equality of means in the preceding sections assumed that the dispersion parameters of the groups are equal. However, this assumption should be checked before proceeding to test the equality of means. Two statistics, namely the likelihood ratio test and the $C(\alpha)$ test are developed. The usual competing hypotheses are

$$H_o : c_1 = c_2 = \ldots = c_k$$

and

$$H_A : \text{at least one inequality among } c_i\text{'s} \qquad (4.6.1)$$

where $m_1, m_2, \ldots, m_k$ are the nuisance parameters under H_o and H_A. Let $y_{ij} \sim NB(m_i, c_i)$ s.t. $i = 1,\ldots,k$; $j = 1,\ldots,n_i$.

Let ℓ be the log-likelihood under the alternative hypothesis.

$$\ell = \sum_i \sum_j \sum_t \ln(1 + c_i t) - \sum_i \frac{n_i}{c} \ln(1 + c_i m_i)$$

$$+ \sum_i y_{i+}(\ln m_i - \ln(1 + c_i m_i)) \qquad (4.6.2)$$

where

$$i = 1,\ldots,k,\ j = 1,\ldots,n_i,\ t = 1,\ldots,y_{ij}-1$$

and $y_{i+} = \sum_j y_{ij}$.

i) Likelihood Ratio Test :

The likelihood ratio test is given by

$$\frac{1}{2}\chi^2_{lc} = \sum_i \sum_j \sum_t \ln \frac{1 + c_i t}{1 + c_c t} - \sum_i n_i \bar{y}_i \ln \frac{1 + \bar{y}_i c_i}{1 + \bar{y}_i c_c}$$

$$+ \sum_i \frac{n_i}{c_c} \ln(1 + c_o \bar{y}_i)$$

$$- \sum_i \frac{n_i}{c_i} \ln(1 + c_i \bar{y}_i) \ . \qquad (4.6.3)$$

Just for this expression, c_i stands for the maximum likelihood estimate of the dispersion parameter of ith group and c_o in this case is the mle of common dispersion parameter under the H_o above. For calculating this statistic ,(k+1) non-linear equations have to be solved iteratively to obtain k c_i's and c_o. Further the statistic has a complicated form.

ii) C(α) test :

Reparametrise c_i's as

$$c_i = c + \phi_i \ , \ i = 1,..,k,$$

with $\phi_k = 0$.

Using the notations described earlier

$$\psi_i(\theta) = \frac{n_i}{c^2} \ln(1 + c m_i) - \frac{1}{c} \sum_j \sum_t \frac{1}{1 + ct} \ , \qquad (4.6.4)$$

$$i = 1,\ldots,k-1,$$

$$\gamma_1(\theta) = \sum_i \frac{n_i}{c^2} \ln(1 + c m_i) - \frac{1}{c} \sum_i \sum_j \sum_t \frac{1}{1 + ct} \ ,$$

$$\gamma_{i+j}(\theta) = \frac{n_j(\bar{y}_j - m_j)}{c(1 + cm_j)}, \qquad j = 1,..,k,$$

$$G_i = n_i c^{-4} \sum_{r=i}^{\infty} \frac{r!(cq_i)^{r+1}}{(r+1)(1+c)\ldots(1+rc)}, \qquad (4.6.5)$$

$$D_{ij}(\theta) = G_i, \qquad 1 \le i = j \le k-1,$$

$$= 0, \qquad \text{otherwise},$$

$$A_{ij}(\theta) = G_i, \qquad 1 \le i < k-1;\ j = 1,$$

$$= 0, \qquad \text{otherwise},$$

$$B_{ii}(\theta) = \sum_i G_i,$$

$$B_{i+j,i+j}(\theta) = \frac{n_j}{m_j(1 + cm_j)}, \qquad (4.6.6)$$

$$j = 1,..,k.$$

For k = 3, the matrix V looks like

$$V = \begin{bmatrix} G_1 & 0 & G_1 & 0 & 0 & 0 \\ 0 & G_2 & G_2 & 0 & 0 & 0 \\ G_1 & G_2 & G_+ & 0 & 0 & 0 \\ 0 & 0 & 0 & H_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & H_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & H_3 \end{bmatrix}$$

Following Collings(1981) and Fisher(1941),

$$D_{ii}(\theta) = E\left[- \frac{\partial^2 \ell}{\partial \phi_i^2} \Big| \Phi = 0 \right] = E\left[- \frac{\partial^2 \ell_i}{\partial c^2} \right]$$

$$= n_i c^{-4} \sum_{r=1}^{\infty} \frac{r!(cq_i)^{r+1}}{(r+1)(1+c)\ldots(1+rc)} = G_i, \qquad (4.6.7)$$

where $q_i = \frac{cm_i}{1 + cm_i}$.

Define

$$S_i(\theta) = \psi_i(\theta) - \beta_{i1}\gamma_1(\theta) - \sum_j \beta_{i,1+j}\gamma_{1+j}(\theta). \qquad (4.6.8)$$

From the pattern of matrix V, we conclude

$$\beta_{i,1+j}(\theta) = 0 \; ; \; j = 1,\ldots,k, \; i = 1,\ldots,k-1,$$

and

$$\beta_{i1}(\theta) = \frac{G_i}{G_+} \qquad i = 1,\ldots,k-1. \qquad (4.6.9)$$

Substituting the values of β_{ij}'s from (4.6.9) in (4.6.8), we get

$$S_i(\theta) = \psi_i(\theta) - \beta_i(\theta)\gamma_1(\theta). \qquad (4.6.10)$$

Following the same steps as in section 4.3 , it follows easily that

$$S_k(\theta) = - (\psi_k(\theta) - \beta_k(\theta)\gamma_1(\theta)) = \sum_{i=1}^{k-1} S_i(\theta).$$

Writing S_i for $S_i(\theta)$ for simplicity and denoting

$S' = (S_1,\ldots,S_{k-1})$ the test statistic has the following form

$$\chi^2_{CC} = S'(D - AB^{-1}A')^{-1}S \ . \qquad (4.6.11)$$

Define a vector $G = (G_1,G_2,\ldots,G_{k-1})$, it follows easily that

$$D - AB^{-1}A' = Diag(G) - \frac{GG'}{G_+} \ .$$

Thus ,

$$\chi^2_{CC} = \sum_{i=1}^{k} \frac{S_i^2}{G_i} \ . \qquad (4.6.12)$$

The C(α) statistics are obtained by replacing the nuisance parameters m_i's and c in (4.6.12) by their $\sqrt{n}$-consistent estimates as follows:

i)MLEs of the parameters m_i and c are $\bar{y}_i$ and c_o respectively. Using these estimates $S_i(\theta)$ reduces to $\psi_i(\theta)$. Hence χ^2_{CC} can be written as

$$\chi^2_{CC1} = \sum_{i=1}^{k} \frac{\psi_i^2(\hat{\theta})}{D_i} \ , \qquad (4.6.13)$$

where

$$\psi_i(\hat{\theta}) = \frac{n_i}{c_1^2} \ln(1 + c_o\bar{y}_i) - \frac{1}{c_o}\sum_j\sum_i \frac{1}{1 + c_o i}$$

ii)The method of moments estimators for m_i and c are $\bar{y}_i$ and c' respectively , where

$$c' = \frac{\sum_i (S_i^2 - \bar{y}_i)}{\sum_i (\bar{y}_i)^2} .$$

In this case $\gamma_1(\theta)$ does not become 0 when c is replaced by c'.Let $\hat{\beta}$, $\hat{\psi}_i$, $\hat{\gamma}_1$ be the values of β_i's, ψ_i's and γ_1 evaluated at $\theta = (\bar{y}_1, \ldots, \bar{y}_k, c')$.

Thus another C(α) test may be constructed as

$$\chi^2_{CC2} = \sum_{i=1}^{k} \frac{(\hat{\psi}_i - \hat{\beta}_i \hat{\gamma}_1)^2}{\hat{G}_i}$$

$$= \sum_{i=1}^{k} \frac{\hat{\psi}_i^2}{\hat{G}_i} - \frac{\hat{\gamma}_1^2}{\hat{G}_+} . \qquad (4.6.14)$$

Expanding $\psi_i(\hat{\theta})$, and $D_i(\hat{\theta})$ about $c_o = 0$, we get

$$\psi_i(\hat{\theta}) = \frac{1}{2} n_i (S_i^2 - \bar{y}_i) + O(c_o^2) ,$$

$$D_i(\hat{\theta}) = \frac{n_i \bar{y}_i^2}{2(1+c_c)(1+c_o \bar{y})^2} + O(c_o^2) .$$

Assuming $c_o = 0$, χ^2_{CC1} reduces to

$$\chi^2_{CC1} = \sum_i \frac{n_i (S_i^2 - \bar{y}_i)^2}{2\bar{y}_i^2} .$$

Thus χ^2_{CC1} will show good power when the values of c_i's are large and significantly different.χ^2_{CC1} has the same form

as the χ^2 statistic for detecting the negative binomial variation in one-way lay-out when the dispersion parameters of groups are not necessarily equal.

4.7 Examples

In this section, two data sets are analysed using the methods presented in the last two chapters.

1. This example is taken from McCaughran & Arnold(1976). Data in table 4.4 presents the counts of embroynic deaths in a control group and two treatment groups along with sample statistics.

Table 4.4 Counts of embryonic deaths in a control and two treatment groups

	frequency		
Numbers of deaths	control group	dose level 1	dose level 2
0	7	5	4
1	2	4	2
2	1	0	3
3	0	1	0
4	0	0	1
$\bar{y}$	0.40	0.70	1.20
s^2	0.44	0.81	1.56
c'	0.25	0.23	0.25

The statistic for testing equality of means , assuming the data follows Poisson distribution, given by χ^2_4 in (3.3.17) is 4.24. Since 4.24 < 5.99 (percentage point of $\chi^2(2)$ at 5 %), we have no evidence to doubt the equality of means

of the groups . The method of moments estimate of c, calculated separately for different groups are 0.25, 0.23, 0.25. This shows the presence of negative binomial variation strongly. From chapter 3, the ststistics T_c (see (3.3.19)) and T_g yield the values 1.203 and 36.49, which do not show the presence of negative binomial variation as

$$1.203 < 1.645 \ (z_{0.95})$$

and $$36.49 < 40.1 \ (\chi^2_{0.95}(27)).$$

The hypothesis of the presence of negative binomial variation is rejected inspite of the large values of dispersion parameters. This anomaly can be explained as follows :

If $y \sim NB(m,c)$, then $E(y) = m$ and $var(y) = m + m^2c$. If m is small the contribution of c towards the variance is reduced. If m= 0.4 and c= 0.25 , then var(y) = 0.4 + 0.04. The variance exceeds the mean only by 0.04. Thus the variation observed in $S^2-\bar{y}$ calculated from several samples will not provide strong evidence of negative binomial variation. Thus, it may not be wise to ignore the over-dispersion in the data. Therefore , we proceed to analyse the data using the negative binomial distribution. The hypothesis of a common dispersion parameter is tested using the statistic χ^2_{CC1} (see 4.6.10) which has the value 0.00462. This indicates a strong evidence regarding the validity of the common dispersion parameter c. The value of the $C(\alpha)$ test statistic for testing equality of means

is 3.007 on 2 d.f. showing that the means of the treatment groups are not significantly different from the control group. This conclusion is in agreement with that of McCaughran & Arnold(1976),who analysed the data using the log-transformation of the data. Note that the value of the test statistic has reduced after taking into account the possible over-dispersion in the data set.

Example 2.This data set is taken from Beall(1939) and refers to counts of adult Colorado potato beatles (Leptinotarsa dicemlineata,say)in each two-foot unit of row in an untreated heavily infested field of potatoes. Each of the sixteen blocks was divided into eight plots ,each two rows wide and 10 ft. long & seperated by single guard rows. For the present purpose , consider the data to consist of sixteen blocks reproduced in the form of a frequency distribution table as shown below in table 4.5. Table 4.6 presents summary statistics for the data in table 4.5.

Table 4.5 Counts of adult potato beatles in a field experiment(condensed from Beall(1939))

No. of Beatles	frequency in blocks							
	1	2	3	4	5	6	7	8
0	6	4	11	30	6	6	3	5
1	8	6	20	22	13	13	9	10
2	11	3	19	19	17	22	11	13
3	9	7	15	21	21	17	14	17
4	10	10	22	16	13	27	16	17
5	11	11	11	13	24	16	10	19
6	7	7	13	6	22	11	18	16
7	13	13	10	6	11	11	11	19
8	9	8	3	4	4	10	11	10
9	13	8	7	4	4	4	10	7
10	10	11	5	1	6	1	7	2
11	7	10	4	2	1	3	6	2
12	4	8	1	0	0	2	3	3
13	9	7	1	0	1	1	4	1
14	3	9	2	0	0	0	2	1
15	1	2	0	0	1	0	2	1
≥16	13	20	0	0	0	0	7	1

No. of Beatles	frequency in blocks							
	9	10	11	12	13	14	15	16
0	3	6	8	18	23	10	39	12
1	7	14	18	21	18	21	47	17
2	11	16	18	32	26	31	23	32
3	15	12	16	29	9	14	25	19
4	22	17	31	19	23	24	7	20
5	10	14	21	9	13	23	1	13
6	17	15	16	7	8	9	2	9
7	9	13	6	3	9	5	0	11
8	16	11	4	3	3	3	0	5
9	13	8	3	2	3	2	0	2
10	10	0	2	0	3	0	0	2
11	3	4	0	0	3	1	0	0
12	1	4	1	0	0	1	0	1
13	4	5	0	1	1	0	0	1
14	2	0	0	0	0	0	0	0
15	0	4	0	0	1	0	0	0
≥16	1	1	0	0	1	0	0	0

Table 4.6

Summary statistics for the data in table 4.2. Estimated means, variances and dispersion parameters for the 16 groups in the above table

groups	means	variances	disp. par.
1	7.51	21.22	0.243
2	8.89	21.99	0.166
3	4.36	10.59	0.328
4	2.99	7.06	0.456
5	4.57	7.62	0.146
6	4.41	7.67	0.168
7	6.56	16.43	0.230
8	5.26	9.76	0.162
9	5.99	10.95	0.138
10	5.51	14.03	0.280
11	3.89	5.49	0.106
12	2.85	4.78	0.236
13	3.63	10.23	0.500
14	3.40	5.21	0.156
15	1.48	1.79	0.143
16	3.59	6.69	0.240

The statistic for testing equality of means under the Poisson assumption attains the value 1791.243(which is significant even at 0.001 level of significance).By looking at the estimates of the means and variances , presence of negative binomial variation is suspected. The test statistic for detecting this negative binomial variation takes the value 36.93 which is highly significant, indicating a strong evidence of the presence

of negative binomial variation. The test-statistic for testing the equality of dispersion parameters c_i's is 3.9169 which is less than 24.996($=\chi^2_{0.05}(15)$), suggesting that the assumption of common c is tenable. The value of the C(α) test for equality of means is 570.917 which is higher than 32.801($=\chi^2_{0.005}(15)$), showing that the means are significantly different among blocks.

CHAPTER 5

ANALYSIS OF ONE-WAY LAY-OUT OF COUNT DATA WITH NEGATIVE BINOMIAL VARIATION WHEN THE DISPERSION PARAMETERS ARE UNEQUAL

5.1 Introduction

In the usual analysis of variance, homoscedascticity plays an important role. In this chapter, we study the analysis of one-way layout of count data with negative binomial variation when the dispersion parameters are unequal. In particular, we develop methods for testing the equality of means in the presence of unequal dispersion parameters and also for testing the equality of means and dispersion parameters simultaneously. Some simulations are conducted for the Behrens-Fisher problem for negative binomial distributions(meaning comparing means of two negative binomial populations when the dispersion parameters are unequal).

5.2 C(α) Statistic for Testing Equality of Means in The Presence of Unequal Dispersion Parameters

Let $y_{ij} \sim NB(m_i, c_i)$, $i = 1,..,k,\ j = 1,..,n_i$.

Reparametrise m_i's as

$m_i = m + \phi_i$, $\quad i = 1,..,k$ and

s.t. $\phi_k = 0$.

The usual competing hypotheses are,

H_o: $\phi_1 = \phi_2 = \ldots = \phi_k$ and

H_A: at least one inequality among ϕ_i's .

Then

$\theta = (m, c_1, c_2, .., c_k)$ is a set of nuisance parameters both under the null and the alternative hypotheses. (5.2.1)

The log-likelihood ℓ of the combined samples is

$$\ell = \sum \ell_i ,$$

where ℓ_i is the log-likelihood for the ith group under alternative hypothesis.

Define the efficient scores ,

$$\psi_i(\theta) = \frac{\partial \ell}{\partial \phi_i}\Big]_{\Phi = 0} , \qquad i = 1,..,k-1,$$

$$\gamma_1(\theta) = \frac{\partial \ell}{\partial m}\Big]_{\Phi = 0}$$

and

$$\gamma_{1+i}(\theta) = \frac{\partial \ell}{\partial c_i}\Big]_{\Phi = 0} , \; i = 1,\ldots,k . \qquad (5.2.2)$$

Now, the log-likelihood ℓ_i for the ith sample is

$$\ell_i = \sum_j \left[\sum_{t=0}^{y_{ij}-1} \ln(1 + ct) \right] - \frac{n_i}{c} \ln(1 + c(m + \phi_i))$$

$$+ y_{i+}(\ln(m + \phi_i) - \ln(1 + c(m + \phi_i))) .$$

By carrying out the usual differentiation wrt ϕ_i's and θ_i's(i.e. m and c_i's) and setting $\Phi = 0$, we obtain the following

$$\psi_i(\theta) = \frac{y_{i+} - n_i m}{m(1+mc_i)} ,$$

$$i = 1,\ldots,k-1 ,$$

$$\gamma_1(\theta) = \sum_{i=1}^{k} \frac{y_{i+} - n_i m}{m(1+mc_i)}$$

and

$$\gamma_{1+i}(\theta) = \sum_j \sum_t \frac{1}{c_i(1+tc_i)} - \frac{n_i}{c_i^2} \ln(1+mc_i) ,$$

$$i = 1,\ldots,k,$$
$$j = 1,\ldots,n_i ,$$
$$t = 0,\ldots,y_{ij} - 1 . \qquad (5.2.3)$$

The information matrix associated with this likelihood is given by

$$V = \begin{bmatrix} A & B \\ B' & D \end{bmatrix} , \qquad (5.2.4)$$

where

A, B, D are matrices of order (k-1)x(k-1), (k-1)x(k+1) and (k+1)x(k+1) respectively with elements

$$A_{ij} = - E \left[\frac{\partial^2 \ell}{\partial\phi_i \partial\phi_j} \Big|_{\Phi=0} \right] ,$$

$$B_{i1} = - E \left[\frac{\partial^2 \ell}{\partial \phi_i \partial m} \Big|_{\Phi = 0} \right] ,$$

$$B_{i,1+j} = - E \left[\frac{\partial^2 \ell}{\partial \phi_i \partial c_j} \Big|_{\Phi = 0} \right]$$

and

$$D_{ij} = - E \left[\frac{\partial^2 \ell}{\partial c_i \partial c_j} \Big|_{\Phi = 0} \right] .$$

After some algebra we find,

$$A_{ij} = \frac{n_i}{m(1+mc_i)} , \qquad 1 \leq i = j \leq k-1,$$

$$= 0 , \qquad \text{otherwise.}$$

$$B_{i1} = \frac{n_i}{m(1+mc_i)} , \qquad 1 \leq i \leq k-1,$$

$$B_{i,1+j} = 0 , \qquad 1 \leq i \leq k-1; \; 1 \leq j \leq k.$$

$$D_{ij} = d_i , \qquad 1 \leq i = j \leq k ,$$

$$= 0 , \qquad \text{otherwise.} \qquad (5.2.5)$$

The exact calculation of d is not necessary in the calculations that follow. Let $a_i = \frac{n_i}{m(1 + mc_i)}$, and define vectors $a' = (a_1, a_2, \ldots, a_{k-1})$, and $d' = (d_1, d_2, \ldots, d_k)$.

In terms of a and d the information matrix of dimension 2kx2k may be written as

$$V = \begin{bmatrix} \text{Diag}(a) & a & 0 \\ a' & 1'a + a_k & 0 \\ 0 & 0 & \text{Diag}(d) \end{bmatrix} . \qquad (5.2.6)$$

Since, asymptotically,

$$(\psi_1, \psi_2, \ldots, \psi_{k-1}, \gamma_1, \ldots, \gamma_{k+1}) \sim MN(0, V) . \qquad (5.2.7)$$

Thus, from (5.2.7), the conditional distribution of $(\psi_1, \psi_2, \ldots, \psi_{k-1} \mid \gamma_1, \ldots, \gamma_{k+1})$ is given by

$$MN(B\Gamma, V_{11.2}) , \qquad (5.2.8)$$

where

$\Gamma' = (\gamma_1, \gamma_2, \ldots, \gamma_{k+1})$,

$\Psi' = (\psi_1, \psi_2, \ldots, \psi_{k-1})$.

B is the matrix of the partial regression co-efficients, of Ψ on Γ, of which the elements are ,

$$\beta_{ij} = \frac{a_i}{1'a + a_k} , \qquad 1 \le i \le k-1, \; j = 1,$$

$$= 0 , \qquad 1 \le i \le k-1, \; 2 \le j \le k+1 , \qquad (5.2.9)$$

(since $V_{ij} = - E \left[\frac{\partial^2 \ell}{\partial \phi_i \partial \theta_j} \right]$ = covariance between ψ_i & γ_j

$$= \quad 0 \ , \qquad \text{for} \qquad 1 \le i \le k-1 \text{ and } 2 \le j \le k+1 \quad).$$

$$V_{11.2} = A - BD^{-1}B' \ .$$

Substituting the values of A,B and D using (5.2.4) and (5.2.6), we get

$$V_{11.2} = \text{Diag}(a) - \frac{a.a'}{1'a + a_k} \ . \qquad (5.2.10)$$

Now construct adjusted scores S_i's as

$$S_i = \psi_i - \beta_{i1}\gamma_1 - \sum_{j=2}^{k+1} \beta_{ij}\gamma_j \ . \qquad (5.2.11)$$

Using the values of elements of B from (5.2.5), S_i in (5.2.11) reduces to

$$S_i = \psi_i - \beta_i\gamma_1 \ , \qquad (5.2.12)$$

where

$$\beta_i = \beta_{i1} = \frac{a_i}{1'a + a_k} = \frac{a_i}{\sum_{t=1}^{k} a_t} \ . \qquad (5.2.13)$$

Thus the C(α) statistic ,

$$\chi_1^2 = (S_1, S_2, .., S_{k-1})' V_{11.2}^{-1} (S_1, S_2, .., S_{k-1}) \sim \chi^2(k-1) \ .$$

After some simplification, we obtain ,

$$\chi_1^2 = \sum_{i=1}^{k-1} \frac{S_i^2}{a_i} + \frac{(\sum_{i=1}^{k-1} S_i)^2}{a_k} \quad . \qquad (5.2.14)$$

From (5.2.3), we find that ,

$$\gamma_1(\theta) - \sum_{i=1}^{k-1} \psi_i(\theta) = \frac{y_{k+} - n_k m}{m(1+mc_i)} = \psi_k(\theta), \qquad (5.2.15)$$

and from (5.2.13) ,

$$\sum_{i=1}^{k-1} \beta_i = \frac{1'a}{1'a + a_k} ,$$

$$= 1 - \frac{a_k}{1'a + a_k} ,$$

$$= 1 - \beta_k \quad . \qquad (5.2.16)$$

Thus using the definition of β_k from (5.2.16) and $\psi_k(\theta)$ from (5.2.15), we have ,

$$\sum_{i=1}^{k-1} S_i = -(\psi_k(\theta) - \beta_k \gamma_1(\theta)) ,$$

$$= - S_k \quad . \qquad (5.2.17)$$

Hence,

$$\chi_1^2 = \sum_{i=1}^{k} \frac{S_i^2}{a_i} \quad . \qquad (5.2.18)$$

This formula involves nuisance parameters $m, c_1, \ldots, c_k$ which are to be replaced by their $\sqrt{n}$ - consistent estimators. It is easy to see that if maximum likelihood estimates $\hat{m}$ and $\hat{c}_i$'s of m and c_i's (i = 1,...,k) are used

then the C(α) statistic is simply ,

$$\chi_2^2 = \frac{n_i (\bar{y}_i - \hat{m})^2}{\hat{m}(1 + \hat{m}\hat{c}_i)} . \qquad (5.2.19)$$

5.3 $\sqrt{n}$-consistent estimators:

Under the null hypothesis there is a common mean m and k different dispersion parameters $c_1,\ldots,c_k$. The estimate of common mean depends on these dispersion parameters.

i)The maximum likelihood estimates :

The mles of m and c_i's are obtained by solving the following system of (k+1) non-linear equations

$$\frac{\partial \ell}{\partial m} = \sum_{i=1}^{k} \frac{n_i (\bar{y}_i - m)}{m(1 + mc_i)} = 0 \text{ and}$$

$$\frac{\partial \ell}{\partial c_i} = \sum_j \sum_t \frac{1}{c_i(1 + c_i t)} - \frac{n_i (\bar{y}_i - m)}{c_i (1 + mc_i)}$$

$$+\frac{n_i}{c_i^2} \ln(1 + mc_i) = 0$$

$$i = 1,\ldots,k . \qquad (5.3.1)$$

These equations are quite complicated as iterative procedures have to be applied in the following way .

Step 0 :

$$c_i^{(0)} = c_i^* , \qquad i = 1,..,k ,$$

where

c_i^* = method of moments estimator of c_i using $\bar{y}_i$,

$$= \frac{S_i^2 - \bar{y}_i}{\bar{y}_i^2} ,$$

$m^{(0)}$= Initial guess for common mean m ,

$= \bar{y}$ (one possible choice) .

Step j :

After obtaining the estimates of $m, c_1, \ldots, c_k$ at (j-1)th step denoted by $m^{(j-1)}, c_1^{(j-1)}, \ldots, c_k^{(j-1)}$. The jth iteration is performed as follows

calculate $c_i^{(j)}$ by solving the ith equation in (5.3.1) using $c_i^{(j-1)}$ and $m^{(j-1)}$ for $i = 1, \ldots, k$,and then solving for $m^{(j)}$. These steps are repeated until the desired accuracy is obtained. This iterative procedure requires solving for c_i at every step using this complicated expression. This method will theoretically yield maximum likelihood estimates of the required parameters. However, this method failed to acheive convergence even in the case of two groups in the simulation study.

ii) Ordinary method of moments estimates : Another set of estimates can be obtained by the method of moments.

By observing the expression in (5.3.1)

$$\sum_i \frac{n_i (\bar{y}_i - m)}{m(1+mc_i)} = 0,$$

the solution should satisfy ,

$$m = \frac{\sum_i w_i \bar{y}_i}{\sum w_i} ,$$

where $w_i = \frac{n_i}{m(1 + mc_i)}$

and $m(1+mc_i)$ is the variance for the ith group under the assumption of common mean. Let s_i^2 be the estimate of the variance , then m^* given by

$$m^* = \frac{\sum_i w_i^* \bar{y}_i}{\sum_i w_i^*} \quad \text{where} \quad w_i^* = \frac{n_i}{s_i^2}$$

satisfies the equation above.

Equating the second central moment to the second sample moment s_i^2, we have

$$s_i^2 = m(1+m\ c_i).$$

Hence

$$c_i = \frac{s_i^2 - m}{m^2}.$$

Using this relation and m^* we obtain

$$c_{i1} = \frac{s_i^2 - m^*}{(m^*)^2}.$$

Clearly, the estimate m^* obtained in this way satisfies

$$\left.\frac{\partial \ell}{\partial m}\right|_{m = m^*} = 0$$

Thus $\gamma_1 = 0 = \sum_{i=1}^{k} \psi_i$.

Using these estimates , the $C(\alpha)$ statistic based on method of moments estimates is

$$\chi_1^2 = \sum_{i=1}^{k} \frac{n_i (\bar{y}_i - m^*)^2}{m^*(1 + m^* c_i)},$$

$$= \sum_{i=1}^{k} \frac{n_i (\bar{y}_i - m^*)^2}{s_i^2} \sim \chi^2(k-1).$$

5.4 Behrens-Fisher Problem for Negative Binomial distribution

5.4.1 The Tests

Testing equality of means of two normal distributions having unequal variances (unknown) is commonly known as the Behrens-Fisher (BF) problem. The analogous BF problem for negative binomial samples may be described as testing equality of means in the presence of unequal dispersion parameters. For this problem we compare the C(α) test statistic, a t-statistic derived from the C(α) statistic but using degrees of freedom similar to that of the Welch's t-statistic for the normal BF problem and a procedure developed by Banerji(1960).

From section 5.2, the C(α) statistic for 2 groups is obtained by substituting k=2 in the formulas derived earlier .The C(α) test is given by

$$\chi_1^2 = \sum_{i=1}^{2} \frac{n_i (\bar{y}_i - m^*)^2}{m^*(1 + m^* c_i)}.$$

As discussed earlier, the C(α) test is based on method of moments estimators m^* and c_i's. So we have

$$m^* = \frac{w_1\bar{y}_1 + w_2\bar{y}_2}{w_1 + w_2} ,$$

$$w_i = \frac{n_i}{m^*(1 + m^* c_i)} , \quad i = 1,2,$$

$$m^*(1 + m^* c_i) = s_i^2 , \quad i = 1,2$$

and

$$w_i = \frac{n_i}{s_i^2} ,$$

where

$$s_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 , \quad i = 1,2.$$

This reduces χ_1^2 to

$$\chi_1^2 = \sum_{i=1}^{2} \frac{n_i (\bar{y}_i - m^*)^2}{s_i^2} .$$

Substituting the expression of m^* in χ_1^2 , we obtain

$$\chi_1^2 = \frac{(\bar{y}_1 - \bar{y}_2)^2}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} .$$

Taking the square-root on both sides , we have

$$T_2 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} .$$

The form of T_2 is similar to the t-statistic with Welch's approximate degree of freedom ν for normal BF problem where

$$\nu = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right]^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}} .$$

When $n_1 = n_2 = NR$, ν is given by

$$\nu = \frac{(NR - 1)(s_1^2 + s_2^2)^2}{s_1^4 + s_2^4} .$$

Banerji (1960) suggested a testing procedure for comparing the means of two normal distributions with unequal variances as follows :

reject the hypothesis of equality of means if

$$(\bar{x}_1 - \bar{x}_2)^2 \geq \frac{t_1^2 s_1^2}{n_1} + \frac{t_1^2 s_2^2}{n_2} ,$$

where

$\bar{x}_1$, $\bar{x}_2$ are sample means of groups 1 and 2 respectively,

s_1^2, s_2^2 are sample variances of groups 1 and 2 respectively,

t_1 and t_2 are upper $\alpha/2$ percentage points with d.f. $(n_1 - 1)$ and $(n_2 - 1)$ respectively.

Under the normality assumption , this procedure has the advantage of keeping the probability of error of the first kind less than or equal to the pre-assigned level of significance α. Following, the form of the test-statistic in (5.4.4), we construct the rejection region for the BF problem for negative binomial distribution as:

reject the hypothesis of equality of means of two negative binomial samples if

$$(\bar{y}_1 - \bar{y}_2)^2 \geq \frac{t_1^2 s_1^2}{n_1} + \frac{t_2^2 s_2^2}{n_2} ,$$

where

$\bar{y}_1$ & $\bar{y}_2$ are sample means and s_1^2 & s_2^2 are sample variances for groups 1 and 2 respectively, of data assumed to have come from negative binomial distribution, and t_1 and t_2 are the percentage points described earlier.

5.4.2 Simulation Study

A small scale simulation study was conducted to compare the size and power of the three procedures

T_1= C(α) test using the χ^2 approximation,

T_2 = C(α) test using the t-distribution with approximate degree of freedom as suggested by Welch,

T_3 = the square of the differences between the group means compared to the weighted sum of the percentage points of t-distributions (Banerji's procedure).

For comparing sizes and powers the values of the means are chosen as

$$m_1 = 5,\ 10,\ 20,\ 50$$

and

$$m_2 = m_1(1+\phi)\ ,$$

where

$$\phi = 0.0,\ 0.2,\ 0.4,\ 0.6,\ 0.8, 1.0.$$

The values of (c_1, c_2) , the pair of dispersion parameters are

(0.01,0.05), (0.01,0.10), (0.01,0.20), (0.05,0.10), (0.05,0.20), (0.10,0.20).

Sample sizes considered are

$$n_1 = n_2 = NR = 10, 20.$$

A sample of size NR is generated from $NB(m_1, c_1)$ and $NB(m_2, c_2)$ separately and the three statistics T_1, T_2, T_3 are calculated based on these samples and compared with respective percentage points. This procedure is repeated 2000 times and number of rejections is counted for the three statistics. The results multiplied by 1000 are given in tables 5.1a and 5.1b. The column under $\phi = 0.0$, gives the empirical levels of significance. For $\alpha = 0.05$, an empirical level of significance will be considered close to α if it falls within the interval [0.04,0.06]. Since the empirical levels of significance are multiplied by 1000 , which changes this interval to [40,60]. Referring to table 5.1a , we observe that for small values of m and small values of c_i's , T_1 holds the level of significance

but fails to maintain for other pairs of dispersion parameters and other values of m. T_2 and T_3 maintain the level of significance quite well. However, the power shown by T_2 is consistently better than that of T_3. Referring to table 5.1b, we observe that T_1 fails to maintain level of thesignificance for all values of m and pairs of dispersion parameters. In this table also, T_2 and T_3 maintain level of significance quite well. The power of T_2 , in this case also, is consistently higher than that of T_3. T_1 is not recommended as it is a liberal test, the empirical levels of significance are greater than 60 in most of the cases.Based on the simulation results, the statistic T_2 is recommended for analysing the BF problem for negative binomial distribution.

Table 5.1a

Empirical value x 10^3: $\alpha = 0.05$; based on 2000 replications. In each block rows 1,2,3 correspond to statistic T_1, T_2, T_3 respectively. NR = 10

$(c_1, c_2) = (0.01, 0.05)$

m \ ϕ	0.0	0.2	0.4	0.6	0.8	1.0
5	52	149	405	680	877	967
	31	98	297	583	813	936
	22	72	251	529	770	912
10	59	231	616	895	984	999
	33	155	516	837	967	995
	28	131	453	798	953	992
20	72	351	811	981	1000	1000
	40	254	730	961	997	1000
	38	213	682	947	996	1000
50	85	496	939	997	1000	1000
	56	385	888	995	1000	1000
	44	350	865	993	1000	1000

$(c_1, c_2) = (0.01, 0.10)$

5	59	137	360	619	822	926
	33	92	260	510	734	872
	25	75	222	463	686	841
10	71	197	531	813	949	989
	40	133	417	721	897	924
	31	109	367	679	879	917
20	76	267	681	918	987	998
	47	180	554	852	967	994
	37	159	521	829	963	994
50	91	351	793	971	997	1000
	59	250	674	929	989	999
	48	221	656	920	989	999

Table 5.1a (contd.)

$(c_1, c_2) = (0.01, 0.20)$

5	73	124	292	513	715	846
	40	72	186	394	589	755
	29	59	169	346	546	716
10	75	150	395	609	836	937
	46	98	287	530	735	864
	39	80	249	494	708	852
20	89	187	478	756	905	973
	52	115	345	618	818	915
	47	103	322	598	806	907
50	106	226	559	816	947	987
	62	129	394	684	861	950
	57	121	382	676	857	950

$(c_1, c_2) = (0.05, 0.10)$

5	63	143	359	610	811	914
	36	95	255	505	721	863
	27	79	218	451	671	829
10	76	198	509	792	932	983
	44	186	400	692	881	916
	31	112	356	650	856	906
20	84	257	636	889	980	996
	50	181	524	820	950	990
	39	156	475	789	934	937
50	89	308	731	942	994	1000
	56	225	621	891	981	997
	44	199	579	868	978	996

Table 5.1a (contd.)

$(c_1, c_2) = (0.05, 0.20)$

5	73	128	288	507	697	836
	42	76	199	398	578	741
	33	62	167	345	541	709
10	77	160	398	640	827	926
	49	101	276	519	720	855
	40	82	239	477	687	837
20	79	186	463	727	881	965
	54	119	332	595	799	897
	46	100	300	570	782	885
50	94	213	525	787	925	982
	61	134	379	656	847	943
	51	122	351	630	836	937

$(c_1, c_2) = (0.10, 0.20)$

5	73	129	289	494	686	826
	44	79	200	388	573	733
	34	61	166	343	525	689
10	80	161	378	620	808	901
	50	99	262	502	697	840
	38	81	231	454	662	816
20	84	182	437	691	860	997
	50	118	315	577	777	884
	41	91	282	535	744	864
50	90	207	484	746	893	973
	58	135	358	622	820	921
	44	112	315	584	796	907

Table 5.1b

Empirical value x 10^3: $\alpha = 0.05$; based on 2000 replications. In each block rows 1,2,3 correspond to statistic T_1, T_2, T_3 respectively. NR = 20

$(c_1, c_2) = (0.01, 0.05)$

5	72	249	678	927	995	1000
	46	204	620	897	992	999
	40	192	603	892	990	998
10	71	409	889	995	1000	1000
	54	341	849	992	1000	1000
	46	313	835	992	1000	1000
20	79	585	975	1000	1000	1000
	58	525	961	1000	1000	1000
	53	507	957	1000	1000	1000
50	78	782	998	1000	1000	1000
	60	720	997	1000	1000	1000
	54	700	997	1000	1000	1000

Table 5.1b (contd.)

$(c_1, c_2) = (0.01, 0.10)$

5	74	230	636	883	985	997
	53	184	569	852	976	996
	45	177	554	839	971	990
10	77	350	814	986	1000	1000
	56	272	758	973	999	1000
	50	260	740	970	999	1000
20	82	470	916	997	1000	1000
	58	394	885	994	1000	1000
	54	374	875	992	1000	1000
50	89	593	970	1000	1000	1000
	56	521	951	1000	1000	1000
	52	507	948	1000	1000	1000

$(c_1, c_2) = (0.01, 0.20)$

5	76	205	541	814	948	986
	58	153	479	757	925	976
	49	141	459	738	915	973
10	83	259	676	923	988	998
	58	201	608	882	977	995
	53	190	592	876	975	995
20	86	325	772	965	995	1000
	60	253	707	944	991	999
	57	240	700	939	989	999
50	85	373	835	978	1000	1000
	58	299	771	966	998	1000
	54	292	766	964	998	1000

Table 5.1b (contd.)

$(c_1, c_2) = (0.05, 0.10)$

5	72	235	625	880	979	998
	55	189	558	837	967	996
	47	179	533	821	964	996
10	82	328	780	967	999	1000
	59	270	724	952	998	1000
	53	253	709	946	997	1000
20	85	431	884	991	1000	1000
	62	353	850	987	1000	1000
	57	336	837	985	1000	1000
50	82	528	937	999	1000	1000
	61	460	914	998	1000	1000
	55	435	907	998	1000	1000

$(c_1, c_2) = (0.05, 0.20)$

5	74	202	530	795	940	945
	56	160	465	735	909	976
	51	144	443	721	898	970
10	82	255	653	902	984	997
	61	203	589	863	972	994
	54	191	575	852	969	993
20	83	303	740	942	993	1000
	62	239	672	916	988	999
	57	229	657	909	986	999
50	82	353	786	967	999	1000
	59	275	727	956	995	1000
	56	264	715	953	995	1000

Table 5.1b (contd.)

$(c_1, c_2) = (0.10, 0.20)$

5	74	204	520	776	929	982
	58	160	455	718	886	917
	52	147	437	701	889	964
10	85	243	629	879	975	997
	60	201	561	834	965	991
	55	188	544	820	960	990
20	84	288	729	937	989	999
	61	278	654	897	980	996
	58	215	636	890	979	995
50	84	323	742	956	993	1000
	64	255	681	933	991	1000
	58	243	665	930	989	1000

5.5 Example

Lawless(1987) presents data on the times to development of mammary tumours for 48 female rats (originally given by Gail et al.(1980)). Out of these 48 female rats, 23 were assigned to group 1(Retinoid) and the remaining 25 were assigned to group 2(control). Table 5.2 shows the number of tumours developed in the rats belonging to groups 1 and 2 along with the sample statistics.

Table 5.2 : Number of Tumors for Rats in Treatment Groups 1 and 2

no. of tumours	groups		combined samples
	1	2	
0	2	0	2
1	7	4	11
2	4	2	6
3	2	3	5
4	2	2	4
5	4	1	5
6	2	2	4
7	0	2	2
8	0	0	0
9	0	3	3
10	0	1	1
11	0	3	3
12	0	1	1
13	0	1	1
$\bar{y}$	2.652	6.04	4.417
s^2	3.618	14.918	12.368
cmm	0.137	0.243	0.408

The values of test statistics T_1 (see 3.3.18) and T_3 (see 3.3.20) for detecting negative binomial variation are 5.26 and 5.57 respectively, which are highly significant. Thus it can be assumed that data in table 5.2 follow negative binomial distribution. The test-statistic χ^2_{CC2} (see 4.6.11) has the value 8.6542, which is highly significant. This suggests that assumption of equality of dispersion parameters is not tenable for this data set.

We employ the three methods discussed earlier to compare the means of the two groups. The ordinary $C(\alpha)$ test statistic has the value 15.22, which is much larger than 3.84 (upper 5 % point). Thus, the hypothesis of equality of means is rejected strongly.

The approximate d.f. calculated using s_1^2 and s_2^2 yields 35.6 and the largest integer contained in it is 35. Since the percentage point of t-distribution corresponding to 0.05 is a monotonically decreasing function of degree of freedom, we have

$$t(30) = 2.042 > t(35).$$

The value of the test statistic is -3.8924, which is highly significant, hence we reject the hypothesis of equality of means of the the groups in table 5.2.

The 97.5 % points of t-distributions with degrees of freedom 22 and 24 are 2.074 and 2.064. The square of the difference between the means is 11.4785 and the weighted sum of squres of the percentage points is 3.2187.

sum of squres of the percentage points is 3.2187.
Banerji's procedure also rejects the hypothesis of equality of means.
All the three test procedures reject the hypothesis of equality of means strongly suggesting that the treatment(Retinoid) is very effective in reducing the development of tumors in the female rats.

5.6 Simultaneous Testing of Equality of Means and Dispersion parameters of Negative Binomial Distributions

So far in our discussion , we were interested in testing homogeneity of either the means or the dispersion parameters under different conditions on the nuisance parameters. In order to test that several negative binomial distributions are identical, we need to test the equality of means and dispersion parameters simultaneously. For this problem a C(α) test is derived.

Let $y_{ij} \sim NB(m_i, c_i)$, $\quad i = 1,..,k,\ j = 1,...,n_i$.
Thus the hypothesis that y_{ij}'s are identically distributed as NB(m,c) where m is the common mean and c is the common dispersion may be written as

$$H_o: m_1 = m_2 = \ldots = m_k \ ; \ c_1 = c_2 = \ldots = c_k$$

and

H_A: at least one inequality among m_i's or c_i's.

Reparametrise

$$m_i = m + \phi_i \ ,$$

$$c_i = c + \tau_i \ , \qquad i = 1,..,k-1 \ ,$$

$$m_k = m \qquad \text{and}$$

$$c_k = c \ .$$

Define $\Phi = (\phi_1,\phi_2,..,\phi_{k-1})$; $T = (\tau_1,\tau_2,..,\tau_{k-1})$.

Let ℓ be the log-likelihood of the data, differentiating ℓ with respect to Φ, T, m, c and substituting 0 for Φ and T in the resulting derivatives, we get,

$$\left.\frac{\partial \ell}{\partial \phi_i}\right|_{(\Phi,T)\,=\,0} = \frac{n_i(\bar{y}_i - m)}{m(1+mc)} \ , \qquad i = 1,..,k-1 \ ,$$

$$\left.\frac{\partial \ell}{\partial m}\right|_{(\Phi,T)\,=\,0} = \sum_i \frac{n_i(\bar{y}_i - m)}{m(1+mc)} \ ,$$

$$\left.\frac{\partial \ell}{\partial \tau_i}\right|_{(\Phi,T)\,=\,0} = \sum_j \sum_t \frac{1}{c(1+ct)} - \frac{n_i(\bar{y}_i - m)}{c(1+mc)} + \frac{n_i}{c^2}\ln(1+mc) \ .$$

$$\left.\frac{\partial \ell}{\partial c}\right|_{(\Phi,T)\,=\,0} = \sum_i \sum_j \sum_t \frac{1}{c(1+ct)} - \sum_i \frac{n_i(\bar{y}_i - m)}{c(1+mc)} + \sum_i \frac{n_i}{c^2}\ln(1+mc),$$

$$j = 1,..,n_i; \ t = 0,..,y_{ij}-1 \ .$$

The information matrix consists of the following block matrices ,

$$A_{ij} = -\,E\left[\frac{\partial^2 \ell}{\partial \phi_i \partial \phi_j}\right] \ ,$$

$$= \frac{n_i}{m(1 + mc)} , \quad 1 \le i = j \le k-1 ,$$

$$= 0 , \quad \text{otherwise} .$$

$$B_{ij} = - E \left[\frac{\partial^2 \ell}{\partial \phi_i \partial \tau_j} \right] = 0 , \quad 1 \le i,j \le k-1 ,$$

$$d_i = - E \left[\frac{\partial^2 \ell}{\partial \phi_i \partial m} \right] ,$$

$$= \frac{n_i}{m(1 + mc)} , \quad 1 \le i \le k-1 ,$$

$$e_i = - E \left[\frac{\partial^2 \ell}{\partial \phi_i \partial c} \right] = 0 , \quad 1 \le i \le k-1 ,$$

$$F_{ij} = - E \left[\frac{\partial^2 \ell}{\partial \tau_i \partial \tau_j} \right] ,$$

$$= f_i , \quad 1 \le i = j \le k-1 ,$$

$$= 0 , \quad \text{otherwise} ,$$

$$g_i = - E \left[\frac{\partial^2 \ell}{\partial \tau_i \partial m} \right] = 0 , \quad 1 \le i \le k-1 ,$$

$$h_i = - E \left[\frac{\partial^2 \ell}{\partial \tau_i \partial c} \right] = f_i , \quad 1 \le i \le k-1 ,$$

$$P_{11} = - E \left[\frac{\partial^2 \ell}{\partial m^2} \right] = \sum_{i=1}^{k} \frac{n_i}{m(1 + mc)} ,$$

$$P_{22} = - E \left[\frac{\partial^2 \ell}{\partial c^2} \right] = \sum_{i=1}^{k} f_i$$

and

$$P_{12} = - E \left[\frac{\partial^2 \ell}{\partial m \partial c} \right] = 0 = P_{21} .$$

Before taking the expectations of the derivatives; (Φ, T) was replaced by 0 reducing m_i to m and c_i to c.Arranging the block matrices properly

$$V = \begin{bmatrix} A & B & d & e \\ B' & F & g & h \\ d' & g' & P_{11} & P_{12} \\ e' & h' & P_{21} & P_{22} \end{bmatrix}$$

Define $d' = (d_1, d_2, \ldots, d_{k-1})$,

$g' = (g_1, g_2, \ldots, g_{k-1})$,

$e' = (e_1, e_2, \ldots, e_{k-1})$,

$h' = (h_1, h_2, \ldots, h_{k-1})$.

The matrix V ,after substituting the values of A,B,e etc., has the form

$$V = \begin{bmatrix} Diag(a) & 0 & d & 0 \\ 0 & Diag(f) & 0 & h \\ d' & 0 & 1'a + a_k & 0 \\ 0 & h' & 0 & 1'f + f_k \end{bmatrix} .$$

Asymptotically ,

$$\left(\frac{\partial \ell}{\partial \Phi}, \frac{\partial \ell}{\partial T}, \frac{\partial \ell}{\partial m}, \frac{\partial \ell}{\partial c} \right) \sim MN(0, V) .$$

Define

$$S_{1i} = \frac{\partial \ell}{\partial \phi_i} - \beta_{i1} \frac{\partial \ell}{\partial m} - \beta_{i2} \frac{\partial \ell}{\partial c} ,$$

$$S_{2i} = \frac{\partial \ell}{\partial \tau_i} - \zeta_{i1}\frac{\partial \ell}{\partial m} - \zeta_{i2}\frac{\partial \ell}{\partial c} \quad ,$$

$$S_1 = (S_{11}, S_{12}, \ldots, S_{1,k-1})$$

and $$S_2 = (S_{21}, S_{22}, \ldots, S_{2,k-1}) \quad .$$

Hence ,

$$(S_1, S_2) V_{11.2}^{-1} (S_1, S_2)' = \chi_4^2 \sim \chi^2 \text{ with } (2k-2) \text{ d.f.}$$

$V_{11.2}$ reduces to a convenient form ,

$$V_{11.2} = \begin{bmatrix} \text{Diag}(a) - \dfrac{aa'}{1'a + a_k} & 0 \\ 0 & \text{Diag}(f) - \dfrac{ff'}{1'f + f_k} \end{bmatrix} .$$

Thus χ_4^2 breaks up into two parts as

$$\chi_4^2 = S_1\left[\text{Diag}(a) - aa'/(1'a + a_k)\right]^{-1} S_1'$$

$$+ S_2\left[\text{Diag}(f) - ff'/(1'f + f_k)\right]^{-1} S_2' \quad ,$$

$$= \chi_1^2 + \chi_2^2 \quad .$$

From the matrix V , it is clear that ,

$\beta_{i2} = \zeta_{i1} = 0$ and $\beta_{i1} = \zeta_{i2} = \dfrac{n_i}{n}$.

After some algebra,it is seen that χ_1^2 is the C(α) test statistic for comparing the means of the negative binomial distributions assuming a common dispersion parameter and χ_2^2 is the C(α) test statistic for testing the equality of dispersion parameters of several negative binomial

distributions assuming common means. This partitioning of χ^2 is similar to the partitioning of the likelihood ratio test statistic for simultaneously comparing the means and variances of several normal distributions.

CHAPTER 6

DETECTION OF OUTLIERS IN POISSON SAMPLES

6.1 Introduction : It has been shown earlier that the departure from Poisson assumption in a data set is determined by the large values of $S^2 - \bar{x}$ or $S^2 / \bar{x}$. It can lead to erroneous conclusion of negative binomial variation due to the presence of one or more outliers. Most of the literature on outliers discusses the methods of detection of outliers in continuous distributions (see Barnett & Lewis, 1978 ; Beckman & Cook , 1983). In this chapter , several methods of detecting one outlier in Poisson samples are discussed.

6.2 Exact test

The usual method of constructing tests involving discrete distributions is to consider conditional tests in order to make the test statistic independent of the nuisance parameters. Doornboos(1966) proposed one method appealing to the conditional distribution theory.

Let $x_1, x_2, \ldots, x_n \sim Poi(\lambda)$, the conditional joint distribution is given by

$$(x_1, \ldots, x_n \mid \sum_{i=1}^{n} x_i = T) \sim Mult(T; 1/n, \ldots, 1/n). \qquad (6.2.1)$$

Then the marginal distribution of an individual x_i is

Binomial(T, 1/n) under (6.2.1). Thus the $\max(x_i) = x_{(n)}$ will be considered an outlier if we could find a cut-off point C such that whenever

$$x_{(n)} \geq C \ , \tag{6.2.2}$$

then $x_{(n)}$ will be declared as an outlier at level of significance α. In statistical framework ,

$$H_0 : x_1, \ldots , x_n \sim \text{Poi}(\lambda) \text{ and}$$

$$H_A : x_1, \ldots, x_{n-1} \sim \text{Poi}(\lambda),$$

$$x_n \sim \text{Poi}(\lambda_1) \tag{6.2.3}$$

where $\lambda_1 > \lambda$ if upper outlier is to be detected and

$\lambda_1 < \lambda$ if lower outlier is to be detected.

The cut-off point is determined by solving

$$P(x_{(n)} \geq C_u \mid \Sigma x_i = T) = \alpha \ , \tag{6.2.4}$$

where C_u is critical point for detecting upper outlier. The distribution of $x_{(n)}$,the largest observation, is intractable in the sense that its percentage points can not be calculated exactly due to the discreteness and complicated expression. The expression (6.2.4) can be written as

$$P(x_1 \geq C_u \text{ or } x_2 \geq C_u \text{ or } \ldots \text{or } x_n \geq C_u) = \alpha, \tag{6.2.5}$$

where

$$(x_1, \ldots , x_n) \sim \text{Mult}(T; 1/n, \ldots ,1/n).$$

Let $A_i = P(x : x_i \geq C_u)$, then (6.2.5) is

$$\alpha = P(\cup A_i) \ ,$$

$$= \sum_i P(A_i) - \sum_{i \neq j} P(A_i \cap A_j) + \ldots \ . \tag{6.2.6}$$

Using Bonferroni's inequality ,approximately

$$\alpha \approx \sum_{i} P(A_i) \quad . \qquad (6.2.7)$$

As $P(A_i)$ is identical for $i = 1,\ldots,n$, hence $P(A_i) = \alpha/n$. Combining all the results above, C_u is obtained as a solution to the expression ,

$P(X \geq c) = \alpha/n$,

$$= \sum_{x=c}^{T} \binom{T}{x} (1/n)^x (1-1/n)^{T-x} \quad . \qquad (6.2.8)$$

Due to the discreteness of the random variable X , we may not acheive the value α/n exactly ; hence we find out the value of C_u such that $P(X \geq C_u) \leq \alpha/n$ .Though this conditional test has the desirable optimal properties , the test is highly conservative as the actual significance level is always smaller. The existing tables (e.g. table XVII, p 316, Barnett & Lewis, 1978) do not offer critical values for all the combinations of sample size, n and sample total, T. Following the same approach as outlined above , the critical value for detecting a lower outlier can be obtained as a solution to the following inequality

$$\sum_{x=c}^{c_l} \binom{T}{x} (1/n)^x (1- 1/n)^{T-x} \leq \alpha/n \quad . \qquad (6.2.9)$$

For large values of n and T , the calculation of critical values C_u and C_l are time consuming.

6.3 Unconditional Tests

In an attempt to develop unconditional tests, two

approaches are considered

i) A statistic based on the generalised likelihood ratio approach (Hawkins , 1980, p 19),

ii) A statistic based on adjusted residuals (Haberman, 1973) .

Defining $\lambda_i = \lambda C$ in the setting of (6.2.3) where

$C > 1$, for upper outlier,

$C < 1$, for lower outlier,

the competing hypotheses are changed to

$$H_O : C = 1 \text{ and}$$
$$H_{AU}: C > 1 \qquad (6.3.1)$$

for the detection of upper outlier

(or $H_{AL}: C < 1$ for the detection of lower outlier) .

The likelihood ratio statistic χ_i^2, for testing that the ith observation is an outlier, is asymptotically distributed as χ^2 with 1 d.f. , has the form

$$\chi_i^2 = 2\{(T - x_i)\ln((T - x_i)/(n-1)) \} . \qquad (6.3.2)$$

This χ_i^2 will have the largest value for the largest x_i, i. e., $x_{(n)}$. Hence $x_{(n)}$ will be considered an outlier for a large value of

$$LRT = \max\{ \chi_i^2 \} . \qquad (6.3.3)$$

A statistic for testing the smallest observation $x_{(1)}$ for a lower outlier can similarly be constructed . Conditional on the sample total, the distribution of LRT may be parameter indepedent but unconditionally the critical values will depend on the Poisson parameter. The

distribution of these statistics have intractable form. The usual index of dispersion test for testing the assumption of data coming from Poisson distribution is

$$D = \sum_i (x_i - \bar{x})^2/\bar{x} ,$$

$$= \sum_{i=1}^{n} e_i^2 . \qquad (6.3.4)$$

D is approximately distributed as χ^2 with (n-1) d.f. The contribution made by ith observation to this test statistic is e_i^2. The observation x_i will be considered an outlier if e_i^2 is significantly large. A test statistic can be constructed based on $e_i = (x_i - \bar{x})/(\bar{x})^{\frac{1}{2}}$ with

$$E(e_i) \approx 0$$

and $var(e_i) \approx (n-1)/n$,

thus the adjusted standardised residual l_i, is given by

$$l_i = \sqrt{n/(n-1)}\ (x_i - \bar{x})/(\bar{x})^{\frac{1}{2}} \qquad (6.3.5)$$

approximately distributed as standard normal variate. In this case also l_i will be the largest for the largest observation $x_{(n)}$.

Thus $x_{(n)}$ is declared an upper outlier for significantly large value of

$$M = l_{(n)} = \max_i \{ l_i \}. \qquad (6.3.6)$$

The statistic for detecting a lower outlier is similarly constructed. The statistic M is also a $C(\alpha)$ test derived using the method of Neyman(1959) and the procedure

discussed in chapter 1.This test is same as the test of an outlier in multinomial sample described in Fuchs & Kennett(1980). This suggests that M -test is independent of parameter conditionally on predetermined sample total, but unconditionally it is parameter dependent. Table 6.1a & 6.1b presents results of a small scale simulation study conducted to study the dependence of percentage points on the parameter , for the following values of λ, n and α

$$n = 10, 20$$

$$\lambda = 5, 10, 15, 25, 50$$

$$\alpha = 0.01, 0.05, 0.01$$

For each pair of n and λ , 15,000 samples were generated and empirical percentage points for LRT and M statistics were obtained respectively.It is clear from the table 6.1a and 6.1b that both these statistics depend on the value of the parameter.

Table 6.1a

Empirical percentiles of LRT for n = 10,20; λ = 5,10,15,25,50; α = 0.10,0.05,0.01 based on 15000 samples from Poisson(λ)

	α = 0.10		α = 0.05		α = 0.01	
λ \ n	10	20	10	20	10	20
5	5.19	6.29	6.36	7.62	9.29	10.65
10	5.27	6.40	6.44	7.63	9.36	10.80
15	5.27	6.41	6.46	7.66	9.33	10.83
25	5.28	6.41	6.47	7.70	9.44	10.87
50	5.30	6.47	6.46	7.74	9.46	10.80

Table 6.1b

Empirical percentiles of M based on 15,000 samples from Poisson(λ). The parameters remain same as in table 6.1a.

	α = 0.10		α = 0.05		α = 0.01	
λ \ n	10	20	10	20	10	20
5	2.58	2.90	2.90	3.23	3.56	3.93
10	2.51	2.81	2.81	3.11	3.43	3.76
15	2.47	2.77	2.76	3.05	3.35	3.69
25	2.44	2.72	2.72	3.00	3.33	3.61
50	2.41	2.68	2.67	2.95	3.25	3.51

6.4 Tests Based on Transformation to Normality :

Conditional and unconditional tests derived so far clearly depend on the sample mean, the estimate of nuisance parameter λ. In this section , statistics based on variance stabilising transformations for Poisson distribution are considered. Barnett & Lewis (1978) suggest using the transformation

$Y_1 = \sqrt{x + 0.25}$, which is approximately distributed as $N(\sqrt{\lambda}, 0.25)$, provided λ is not small, and then use N σ test with $\sigma = 0.5$ on the transformed data. Another transformation due to Anscombe (1950) is

$$Y_2 = \sqrt{x + 0.375} \; . \qquad (6.4.1)$$

The variances are given by

$$Var(Y_2) = 1/4 + \lambda^{-2}/64 + \ldots$$

and

$$Var(Y_1) = 1/4 + \lambda^{-1}/32 + \ldots \; .$$

Clearly , $var(Y_2)$ is closer to 1/4 than $var(y_1)$. Hence in this section only the transformation Y_2 is used. Given the sample $x_1, \ldots, x_n$ let the transformed and ordered sample using (6.4.1) be $y_1 \leq \ldots \leq y_n$.

The first statistic

$$T = 2(y_n - y_{n-1}) \qquad (6.4.2)$$

(for upper outlier)

$$T^* = 2(y_2 - y_1) \qquad (6.4.3)$$

(for lower outlier)

under the assumption that $E(y) = \mu$, unknown but var(y) is known Var(y) is 1/4 when λ is large. For moderate λ , T (T^*) is likely to depend on the Poisson parameter. Critical values of the statistics T and T^* based on 15,000 samples from Poisson(λ) for selected values of the parameter λ , sample size n, level of significance α are given in table 6.2a and 6.2b respectively. It is clear from the tables that the percentage points depend on the parameter and the sample size .The reason for this behaviour is mathematically very cumbersome , hence only a heuristic explanation is given.

Let $S_n = y_n - y_{n-1}$, the difference between the nth and (n-1)th order statistics . Under the assumption of observations coming from the same population , the distribution of S_n will converge to a distribution of a degenerate r.v. taking the value 0 with prob 1 as n tends to infinity. With large λ , the S_n will have more non-zero values than in the case of small λ. This is further enhanced by the fact that due to the discreteness of the r.v. ,S_n will actually assume the value 0 frequently. For this reason the critical values obtained by simulation donot agree with those obtained from normal case even for large λ . A more reasonable approach, perhaps, would be to construct a test statistic by assuming that both E(y) and var(y) are unknown. A number of test statistics are available for this situation e.g. the maximum (minimum) normed or equivalently the Grubbs type statistics and the

normed or equivalently the Grubbs type statistics and the Dixon type statistics discussed in Grubbs(1950, 1969), Dixon(1950) etc. A small scale simulation study showed a great variation in the empirical percentage points of Grubbs statistics in comparison to reasonable parameter independence observed in Dixon type statistics.

Dixon type statistics are

$$TD = (y_n - y_{n-1})/(y_n - y_1) \qquad (6.4.4)$$

(for upper outlier)

$$TD^* = (y_2 - y_1)/(y_n - y_1) \qquad (6.4.5)$$

(for lower outlier)

Critical values of TD and TD^* based on 15,000 samples from Poisson(λ) for selected values of λ, α and n are given in table 6.2b. These critical values show some dependence on the Poisson parameter, however, TD shows better parameter independence and better agreement with critical values obtained using normal samples.

Table 6.2a

Critical values of T and TD for n = 5, 10, 20, 30, 50, 100 and λ = 5, 10, 25, 50, 100 for α = 0.10, 0.05, 0.01 based on 15,000 samples from Poisson(λ) distribution. The last line for each n gives critical values when the samples have been drawn from N(0,1) distribution.

n	α / λ	0.10 T	0.10 TD	0.05 T	0.05 TD	0.01 T	0.01 TD
5	5	1.39	0.54	1.60	0.64	2.26	0.81
	10	1.40	0.54	1.66	0.64	2.28	0.78
	15	1.37	0.54	1.70	0.64	2.29	0.79
	25	1.39	0.54	1.71	0.64	2.31	0.78
	50	1.40	0.54	1.73	0.63	2.32	0.77
	100	1.38	0.54	1.69	0.63	2.28	0.77
	N	1.43	0.55	1.76	0.64	2.36	0.78
10	5	1.07	0.35	1.31	0.41	1.80	0.51
	10	1.10	0.34	1.35	0.40	1.83	0.52
	15	1.12	0.34	1.37	0.41	1.89	0.52
	25	1.15	0.34	1.41	0.40	1.93	0.52
	50	1.16	0.34	1.41	0.40	1.94	0.52
	100	1.11	0.33	1.36	0.39	1.85	0.51
	N	1.19	0.35	1.45	0.41	2.01	0.52
20	5	0.96	0.24	1.19	0.29	1.60	0.37
	10	0.96	0.24	1.22	0.29	1.66	0.39
	15	1.00	0.24	1.22	0.29	1.71	0.38
	25	1.00	0.24	1.20	0.29	1.71	0.38
	50	1.00	0.24	1.24	0.29	1.73	0.38
	100	0.93	0.23	1.18	0.28	1.66	0.37
	N	1.03	0.25	1.28	0.30	1.82	0.40
30	5	0.91	0.19	1.01	0.24	1.53	0.32
	10	0.93	0.20	1.12	0.25	1.58	0.33
	15	0.84	0.21	1.07	0.25	1.59	0.33
	25	0.86	0.21	1.14	0.25	1.60	0.33
	50	0.88	0.21	1.13	0.25	1.60	0.34
	100	0.89	0.20	1.09	0.24	1.54	0.32
	N	0.94	0.21	1.18	0.26	1.71	0.35

(contd.)

Table 6.2a (contd)

	5	0.87	0.17	0.96	0.21	1.46	0.29
	10	0.75	0.17	0.99	0.21	1.44	0.30
	15	0.81	0.18	1.02	0.21	1.53	0.31
	25	0.82	0.18	1.02	0.21	1.56	0.30
50	50	0.86	0.18	1.09	0.22	1.57	0.31
	100	0.81	0.17	1.00	0.17	1.44	0.29
	N	0.89	0.18	1.13	0.22	1.64	0.31
	5	0.62	0.14	0.87	0.18	1.35	0.26
	10	0.71	0.14	0.93	0.18	1.37	0.27
	15	0.78	0.14	0.97	0.18	1.38	0.24
	25	0.79	0.15	0.97	0.18	1.43	0.25
100	50	0.74	0.15	0.97	0.18	1.45	0.25
	100	0.72	0.14	0.90	0.17	1.32	0.23
	N	0.82	0.15	1.03	0.19	1.48	0.25

Table 6.2b

Critical values of T^* and TD^* based on 15,000 samples from Poisson distribution for the parameters defined in Table 6.2a.

n	α / λ	$T^{*0.10}$	TD^*	$T^{*0.05}$	TD^*	$T^{*0.01}$	TD^*
5	5	1.55	0.58	1.88	0.69	2.96	0.84
	10	1.49	0.59	1.94	0.67	2.68	0.82
	15	1.53	0.59	1.88	0.66	2.66	0.81
	25	1.50	0.57	1.83	0.66	2.56	0.81
	50	1.49	0.57	1.83	0.65	2.53	0.79
	100	1.53	0.57	1.88	0.66	2.63	0.81
	N	1.46	0.56	1.76	0.64	2.45	0.79
10	5	1.33	0.39	1.84	0.47	2.45	0.58
	10	1.25	0.38	1.60	0.44	2.35	0.56
	15	1.25	0.37	1.60	0.43	2.26	0.58
	25	1.22	0.36	1.54	0.42	2.15	0.54
	50	1.23	0.36	1.51	0.42	2.12	0.53
	100	1.28	0.36	1.57	0.43	2.24	0.55
	N	1.17	0.34	1.44	0.41	2.01	0.53
20	5	1.33	0.32	1.86	0.38	2.45	0.47
	10	1.25	0.28	1.38	0.33	2.11	0.44
	15	1.07	0.28	1.39	0.33	2.11	0.42
	25	1.12	0.27	1.40	0.32	1.99	0.42
	50	1.11	0.26	1.35	0.32	2.04	0.42
	100	1.12	0.27	1.44	0.32	2.04	0.42
	N	1.03	0.25	1.28	0.30	1.80	0.39
30	5	1.12	0.23	1.86	0.34	1.86	0.41
	10	1.10	0.24	1.38	0.30	1.97	0.39
	15	1.07	0.24	1.39	0.29	1.94	0.37
	25	1.02	0.23	1.30	0.28	1.83	0.36
	50	1.00	0.22	1.28	0.27	1.84	0.35
	100	1.09	0.24	1.35	0.28	1.93	0.37
	N	0.96	0.22	1.19	0.26	1.71	0.33

(contd.)

Table 6.2b (contd.)

50	5	1.12	0.21	1.12	0.25	1.86	0.36
	10	1.10	0.22	1.38	0.27	1.97	0.34
	15	1.07	0.21	1.25	0.25	1.94	0.35
	25	1.02	0.20	1.26	0.24	1.79	0.32
	50	0.97	0.19	1.19	0.23	1.74	0.31
	100	1.02	0.21'	1.27	0.25	1.87	0.33
	N	0.90	0.19	1.13	0.23	1.62	0.31
100	5	1.12	0.20	1.12	0.21	1.86	0.32
	10	1.10	0.20	1.33	0.21	1.84	0.31
	15	0.87	0.16	1.25	0.21	1.76	0.29
	25	0.84	0.16	1.07	0.20	1.65	0.27
	50	0.86	0.16	1.06	0.19	1.60	0.27
	100	0.93	0.17	1.18	0.21	1.76	0.28
	N	0.82	0.15	1.02	0.19	1.50	0.25

In order to compare the powers of T and TD (statistics for detecting an upper outlier), a simulation study was conducted for

$$n = 10, 20,$$

$$\lambda = 10, 50 \text{ and}$$

$$\alpha = 0.10, 0.05, 0.01.$$

For each sample the largest observation $y_{(n)}$ was replaced by $cy_{(n)}$ for c = 1.0, 1.1, 1.2, 1.4, 1.8 and tested whether the new $y_{(n)}$ is detected as an outlier. Normal critical values are used for 10,000 samples generated to count the number of detections (presented in table 6.3). From table 6.3, it is observed that TD holds nominal significance level better than T. However, for larger values of c power of T is in general better than that of TD.

Table 6.3

Monte Carlo estimates of Power (in percent) of the test statistics (test stats) T and TD based on 10,000 samples for n = 10,20; λ = 10,50; α = 0.10,0.05,0.01. The largest value $x_{(n)}$ in each sample is increased to $cx_{(n)}$ for c = 1.0, 1.1, 1.2, 1.4, 1.8.

n	λ	α %	test stat	c 1.0	1.1	1.2	1.4	1.8
		10	T	8.56	18.71	50.16	99.97	100.00
			TD	8.77	19.97	37.94	79.49	99.91
	10	5	T	3.87	10.50	30.01	83.34	100.00
			TD	3.96	11.31	21.26	56.19	97.21
		1	T	0.58	2.28	8.94	31.62	100.00
			TD	0.88	2.29	5.28	18.19	64.30
10		10	T	8.94	45.90	100.00	100.00	100.00
			TD	9.17	39.43	86.56	99.28	100.00
	50	5	T	4.46	27.42	94.26	100.00	100.00
			TD	4.67	23.01	62.07	99.28	100.00
		1	T	0.85	8.64	40.02	100.00	100.00
			TD	0.88	5.55	20.58	75.74	99.92
		10	T	6.69	24.41	48.79	100.00	100.00
			TD	9.05	24.34	53.39	97.74	100.00
	10	5	T	3.47	12.65	45.49	99.93	100.00
			TD	4.11	14.25	32.56	87.81	100.00
		1	T	0.58	3.14	6.26	45.76	100.00
			TD	0.46	2.19	6.49	26.80	94.39
20		10	T	8.40	59.35	100.00	100.00	100.00
			TD	9.24	55.31	98.99	100.00	100.00
	50	5	T	4.01	38.42	100.00	100.00	100.00
			TD	4.47	31.19	88.95	100.00	100.00
		1	T	0.72	8.00	54.22	100.00	100.00
			TD	0.58	6.64	32.77	97.80	100.00

This can be explained by observing the rate of increase in the test statistics by a proportionate increase in the value of the largest observation. Replace $y_{(n)}$ by $(1+\varepsilon)y_{(n)}$ in T and TD respectively,

$$T_{\varepsilon} = 2(y_{(n)} - y_{(n-1)}) + 2\varepsilon y_{(n)}$$

$$TD_{\varepsilon} = \frac{y_{(n)} - y_{(n-1)} + \varepsilon y_{(n)}}{y_{(n)} - y_{(1)} + \varepsilon y_{(n)}} .$$

While T_{ε} and TD_{ε} are both increasing functions of ε, but T_{ε} increases faster than TD_{ε}. Thus for large values of y_n , T should be used and for moderate values TD should be used.

6.5. CONCLUSION :

A conditional test given by Doornboos(1966) has desirable optimality properties though conservative, still it cannot be used readily as the percentage points are not easily obtainable. Two unconditional tests, the likelihood ratio test and the adjusted standardised residual are highly dependent on parameter. Dixon type statistics on the transformed data T and TD have performed better in terms of parameter independence. As shown earlier power of T is better than that of TD

for detecting a large observation as an outlier, however, TD holds level of significance level better than T. This falls in line with the suggestion of Barnett & Lewis(1978).Although methods have been studied for detecting both an upper and lower outlier, but for practical reasons one would be interested in detecting an upper outlier from Poisson data. More specifically, we are interested in the alternative of extra-poisson variation, which provides more weight to the tail probabilities than given by Poisson distn. Hence a large value might either be an outlier or an indicator contributing to the evidence of extra-poisson variation.

Example: As an example, consider data on number of fossils in 1 m^2(Derman et. al. 1973). There were

0, 1, 2, 3, 4 fossils in

16, 9, 3, 1, 1 m^2 respectively. We wish to test if 4 is an outlier. The observed value of T and TD are 0.5091 and 0.1721 respectively, by comparing these with the percentage points from table II , it is not declared as an outlier at 5 % level of significance. However ,if 4 is changed to 5;the new values of T and TD are 0.9626 and 0.2821 respectively. By comparing these with the percentage points sutiably, we could treat 5 as an outlier at 5 % level of significance.

CHAPTER 7

CONCLUSIONS

7.1 Contributions

In analysing one-way layout of count data , the underlying distribution is often assumed to be Poisson. This assumption fails to take into account the overdispersion or extra-poisson variation. Among several distributions available in the literature, negative binomial distribution has been used to model the count data with overdispersion . Based on the ease of calculation of percentage point and the comparable power performance the range-justified statistic is recommended for detection of negative binomial variation.

For comparing the means of groups presented as one-way layout of data asuuming the underlying distribution to be negative binomial, first the validity of assumption of common dispersion parameter should be checked. If the assumption of common dispersion parameter is tenable, use $C(\alpha)$ test with maximum likelihood estimates of nuisance parameters for testing the equality of means.

However, if the assumption of common dispersion parameter is not tenaable, use Welch's approximate degree of freedom formula for testing the equality of means.

A rough method of checking the presence of one outlier in the groups is testing for an outlier assuming the

observations follow Poisson distribution.

7.2 Recommendation

In deriving the $C(\alpha)$ statistics, the central limit theorem for convergence of estimated scores was assumed, without taking into account the skewness or kurtosis of the parent distributions. The effect of skewness and kurtosis of a distribution on the convergence of estimated scores to normality should be examined in details.

Overdispersion in the data was modelled by negative binomial distribution. Other distributions should also be used as the underlying distribution as an alternative to Poisson distribution for comparing the means of the groups.

The analysis of one-way layout of the count data, when the dispersion parameters are unequal, was studied using the simulation study for 2 groups. A comprehensive study is suggested to study the effect of unequal dipersion parameters on the inferences concerning the means.

REFERENCES

Anderson, T.W. (1984). *Introduction to Multivariate Analysis*, 2nd Edn. Wiley, New York.

Anscombe, F.J. (1948). The transformation of Poisson, binomial and negative binomial data. *Biometrika* 38, 246-54.

Anscombe, F.J. (1949). The analysis of insect counts based on the negative binomial distribution. *Biometrics* 5, 165-73.

Anscombe, F.J. (1950). Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika* 37, 358-82.

Banerji, S.K. (1960). Approximate confidence interval for linear functions of k populations when the population variances are not equal. *Sankhya* 22, 357-58.

Barnett, V. and Lewis, T. (1978). *Outliers in Statistical Data.* Wiley, New York.

Bartlett, M.S. (1936). Some notes on insecticide tests in the laboratory and in the field. *J. Roy. Stat. Soc., supp.* 3, 185-94.

Bartoo, J.B. and Puri, P.S. (1967). On optimal asymptotic tests of composite hypothesis. *Ann. Math. Stat.* 38, 1845-52.

Beall, G. (1939). Methods of estimating the population of insects in a field. *Biometrika* 30, 422-39.

Beckman, R.J. and Cook, R.D. (1983). Outlier . . . . s. *Technometrics* 25, 119-63.

Bickel, P.J. and Docksum, K.A. (1977). *Mathematical Statistics : Basic Ideas and Selected Topics.* Holden-Day Inc. San Francisco, CA.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975) . *Discrete Multivariate Analysis: Theory and Practice* . Cambridge, MA: The Massachusetts Institute of Technology .

Bliss, C.I. and Fisher, R.A. (1953). Fitting the negative binomial distribution to biological data. *Biometrics* 9, 176-200.

Bliss, C.I. and Owen, A.R.G. (1958). Negative binomial distributions with a common k. *Biometrika* 45, 37-58.

Bühler, W.J. and Puri P.S. (1966). On optimal asymptotic tests of composite hypothesis with several constraints. *Zeit Wahrscheinlichkeitstheorie* 5, 71-88.

Collings, B.J. (1981). *The negative binomial distribution: an alternative to the Poisson.* Ph.D. thesis, The University of North Carolina at Chapel Hill, NC.

Collings, B.J. and Margolin, B.H. (1985). Testing goodness of fit for the Poisson assumption when observations are not identically distributed. *J. Amer. Stat. Assoc.*, 80, 411-8.

Chant, D. (1974). On asymptotic tests of composite hypotheses in nonstandard conditions. *Biometrika* 61, 291-98.

Cramer, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press.

Derman, C., Gleser, L.J. and Olkin, I. (1973). *A Guide to Probability Theory and Applications.* Holt, Rinehart and Winston.

Dixon, W.J. (1950). Analysis of extreme values. *Ann. Math. Stat.* 21, 488-506.

Doornboos, R. (1966). *Slippage Tests.* Mathematical Centre Tracts, No. 15, Mathematisch Centrum, Amsterdam.

Engel, J. (1984). Models for response data showing extra-Poisson variation. *Statist. Neerlandica,* 38, 159-167.

Fisher, R.A. (1941). The negative binomial distribution. *Annals of Eugenics*, London 11, 182-87.

Fuchs, C. and Kennett, R. (1980). A test for detecting outlying cells in the multinomial distribution and two-way contingency tables. *J. Amer. Stat. Assoc.* 75, 395-98.

Gail, M.H., Santner, T.J., and Brown, C.C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumour. *Biometrics* 36, 225-31.

Grubbs, F.E. (1950). Sample criteria for testing outlying observations. *Ann. Math. Stat.* 21, 27-58.

Grubbs, F.E. (1969). Procedures for detecting outlying observations in samples. *Technometrics* 11, 1–21.

Haberman, S.J. (1973). The analysis of residuals in cross-classified tables. *Biometrics* 29, 205–20.

Haberman, S.J. (1974). *The Analysis of Frequency Data.* The University of Chicago Press.

Hawkins, D.M. (1980). *Identification of Outliers.* London: chapman and Hall.

Hinz, P. and Gurland, J. (1968). A method of analysing untransformed data from the negative binomial and other contagious distributions. *Biometrika* 55, 163–70.

Hinz, P. and Gurland, J. (1970). Atest of fit for the negative binomial and other contagious distributions. *J. Amer. Stat. Assoc.* 65, 887–903.

International Mathematical and Statistical Libraries (1960). " *IMSL Manual* ", Vol II.

Johnson, N.L. and Kotz, S. (1969). *Discrete Distributions.* Wiley, New York.

Kendall, M.G. and Stuart, A. (1979). *The Advanced Theory of Statistics*, Vol. 2. Charles Griffin & Co. Ltd., London.

Kocherlakota, K. and Kocherlakota, S. (1985). On some tests for independence in nonnormal situations: Neyman's C(α) test. *Commun. Statist.- Theor. Meth.* 14, 1453-70.

Lawless, J.F. (1987). Regression methods for Poisson process data. *J. Amer. Stat. Assoc.* 82, 808-15.

Light, R.J. and Margolin, B.H. (1971). An analysis of variance for categorical data. *J. Amer. Stat. Assoc.* 66, 534-44.

Margolin, B.H. (1985). Statistical studies in genetic Toxicology: A perspective from the U.S. National Toxicology Program. *Environ. Health Persp.* 63, 187-94.

McCaughran, D.A. and Arnold, D.W. (1976). Statistical models for numbers of implantation sites and embryonic deaths in mice. *Toxicol. Appl. Pharmacol.* 38, 325-33.

Moran, P.A.P. (1970). On asymptotically optimal tests of composite hypotheses. *Biometrika* 57, 45-55.

Moran, P.A.P. (1973). Asymptotic properties of homogeneity tests. *Biometrika* 60, 79-85.

Neyman, J. (1959). Optimal asymptotic tests for composite hypotheses. In *Probability and Statistics*, Ed. U. Grenander, 213-34. Wiley, New York.

Neyman, J. and Scott, E.L. (1966). On the use of C(α) optimal tests of composite hypotheses. *Bull. Inst. Int. Statist.* 41, 477-97.

Paul, S.R., Liang, K.Y. and Self S.G. (1989). On testing departure from the binomial and multinomial assumptions. *Biometrics* 45, 231-36.

Paul, S.R. and Plackett, R.L. (1978). Inference sensitivity for Poisson mixtures. *Biometrika* 65, 591-602.

Plackett, R.L. (1981). *The Analysis of Categorical Data*. Kent: Griffin.

Pothoff, R.F. and Whittinghill, M. (1966). Testing for homogeneity II: The Poisson Distribution. *Biometrika* 53, 183-90.

Rao, C.R. (1948). Large sample tests of statitical hypotheses concerning several parameters with application to problems of estimation. *Proc. Camb. Phil. Soc.* 44, 50-57.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications.* Wiley, New York.

Self, S.G. and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Stat. Assoc.* 82, 605-10.

Subrahmaniam, K. (1966). A test for "intrinsic correlation " in the theory of accident proneness. *J. Roy. Statist. Soc. B*, 28, 180-89.

Tarone, R.E. (1979). Testing the goodness of fit of the binomial distribution. *Biometrika* 66, 585-90.

Tarone, R.E. (1985). On heterogeneity tests based on efficient scores. *Biometrika* 72, 91-95.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* 54, 426-82.

Wilson, L.J., Folks, J.L. and Young, J.H. (1984). Multistage estimation compared with fixed-sample-size estimation of the negative binomial parameter k. *Biometrics* 40, 109-117.

Zelterman, D. and Chen, C. (1988). Homogeneity tests against central-mixture alternatives. *J. Amer. Stat. Assoc.* 88, 179-82.

VITA AUCTORIS

The author was born at Jamalpur, Bihar, India on July 10, 1961.

He received his B.Stat.(Hons.) from the Indian Statistical Institute in 1981.

He received M.Stat. with specialisation in Statistical Quality Control and Operations Research (SQC & OR) from the Indian Statistical Institute in 1982.

He received his advanced diploma in SQC & OR from the Indian Statistical Institute in 1983.

He worked as a Statistical Consultant in Bureau of Industrial Costs and Prices.

He worked as a fellow in the specialist development program of SQC & OR dept. in Indian Statistical Institute.